NEO

ISSUE I

UNDERGRADUATE JOURNAL OF PHILOSOPHY

NYU PHILOSOPHY FORUM

NEO Undergraduate Journal of Philosophy

Table of Contents

2023 Annual Undergraduate Philosophy Conference

Nietzsche's Critique of Hegel's Philosophy of History

Faateh Ali — *University of Toronto*

Introduction

This essay explores whether history is progressing. In *On the Genealogy of Morality* (1887), Friedrich Nietzsche makes the case that history is not progressing because major historical shifts have just been the outcome of different power dynamics whereby the nature and meaning of social and political institutions have been continually reinterpreted and transformed. In contrast, in the *Introduction to the Philosophy of History* (1837), Georg Wilhelm Friedrich Hegel argues that history is, in fact, progressing insofar as civilizations have become increasingly aware of themselves as free and have, as such, self-consciously developed their institutions so as to realize more 'freedom'. In what follows, I offer critical accounts of both positions and make the case that, ultimately, Nietzsche's is more convincing. That is, although Hegel offers a compelling account for the idea that history is progressively unfolding, I will argue that Nietzsche's strategy of undermining any normative framework by which we might gauge progress in history undermines Hegel's entire attempt at evaluating history.

My paper will proceed in three major steps. First, I shall outline Nietzsche's critique of those that impose a single meaning into their analysis of history, and then show why Nietzsche's 'genealogical' method avoids this error. I will then unpack how genealogy reveals history to not be progressing by exposing major historical changes as nothing but the result of the conflicting interpretive worldviews. Second, I will shift my focus to Hegel in order to consider a serious

objection to Nietzsche's position; I will make the case that despite Nietzschean anxieties, Hegel may be able to provide us with the resources to maintain that, once construed correctly, history is progressing. I will proceed by explicating Hegel's conception of 'Spirit' as capturing the degree to which a civilization is aware of (and has realized) themselves as 'free', and that history has progressed insofar as Spirit has. Finally, I will return to Nietzsche and employ his genealogy in order to criticize Hegel's account by asking the following question: why should we accept that Hegel's standard for evaluating history offers "the" stance for correctly construing history? This criticism, I will show, is both convincing and reveals why Hegel's approach to history cannot do all the philosophical work Hegel thinks it can.

Nietzsche's Genealogical Method

Appreciating Nietzsche's genealogical approach to historical analysis requires us to first consider his criticism of the "English psychologists" and their attempt at a natural account of the history and origins of morality. These psychologists, Nietzsche writes, speculate that moral actions came to be valued and eventually institutionalized in our cultures because, initially, such behaviour benefited those to whom they were shown. Thus, the psychologists locate the origins of morality in its "usefulness" to the recipient of moral actions (GM 1:2). Nietzsche, however, makes the case that morality's original meaning has undergone a series of transformations and reinterpretations in history such that its contemporary instantiation – the interpretation of it that forms the basis of the English psychologists' attempt to discover morality's origin – is profoundly different. As such, Nietzsche criticizes the English psychologists for conflating morality's contemporary and original meaning: "[they] think in a way that is… unhistorical… usefulness was not the [original] concern! … usefulness is alien and inappropriate" (GM 1:2). That is, Nietzsche criticizes the psychologists

for “innocently plac[ing] the purpose at the start” (GM 2:12), i.e., they wrongfully posit today’s meaning as the cause for morality’s emergence, and then proceed to fallaciously trace the history of morality as the series of ways in which morality has been useful, which falsely reads a single meaning of morality in and through its past.

To avoid this error and to properly get morality’s history into view, Nietzsche takes up a genealogical approach to historical analysis. To offer a ‘genealogy’ of, say, a value, institution, or concept is to unearth the conditions under which it first emerged and the various meanings it has taken on since then such that (1) we do not read a single meaning into its past and (2) we can better understand how we are situated in our present-day relation to it so as to correctly construe the possibilities it circumscribes for us.

Nietzsche’s genealogy outlines the series of (re)interpretations morality took on leading up to the birth of Christianity in Judea, and ultimately reveals all such interpretations as just posited by and for some group’s arbitrary interest in establishing power. Nietzsche explains that Judea originally gauged “good” and “bad” by the degree one could engage in ascetic behavior (e.g., fasting and celibacy). However, Nietzsche makes the case that this standard was injected into the collective’s imagination by the ‘priests of Judea’ who were already masters of asceticism. As such, Nietzsche thinks the priests were perceived as morally superior and so were able to establish themselves as the ruling class. Nietzsche proceeds to explain that Judea was later invaded, and that these priests were overthrown by (physically superior) Romans, and as such, the Romans reinterpreted morality to their advantage by positing strength as “the” standard for distinguishing “good” from “bad”. The priests eventually took back power, Nietzsche writes, but because the Romans were physically superior, the priests could not have overthrown the Romans in the same way they did the priests. Thus, Nietzsche argues that the priests took back power from the romans

by reinterpreting morality to their advantage once again such that 'good' was now weakness while 'bad' was strength: "only those who suffer are good… whereas you… the powerful, you are… eternally wretched, cursed and damned!" (GM 1:10).

Nietzsche: History is Not Progressing

By unearthing morality's genealogy, Nietzsche makes the case that our conception of morality is not universal and unchanging, but contingent on factors of our era and the way in which we have inherited it from those who have come before us. That is, genealogy shockingly exposes our understanding of morality to denote a series of errors in our most implicit modes of judging, receiving, and navigating the world. Indeed, not only does genealogy reveal that morality has undergone a series of (re)interpretations throughout history, but also that the vehicle that moves history itself, and so the force that has moved morality from each of its past instantiations, is a power dynamic. As such, morality, genealogy reveals, is nothing over and above a series of arbitrary values that different power-hungry groups have injected into the collective's imagination as "the" universal standard for evaluating *who* is "good" and "bad". Genealogy also reveals that such groups posit these values as such because they function to reflect, justify, and institutionalize the positing-group's worldview so that we come to value that group in a privileged light (i.e., as good) such that the group is well-positioned to establish and maintain power.

Furthermore, genealogy also exposes morality so as to draw the following philosophical conclusion: that *history itself is not progressing*. As Nietzsche makes plain, "the development of a thing… is not its progression" (GM 2:12). Indeed, genealogy exposes history as having changed, not progressively, but in the same way as we understand the evolution of a species. That is, in the

same way we recognize that a species has not "improved", but merely adapted to their present-day environment when they change (i.e., adapt), Nietzsche's genealogy reveals that the changes in the nature and meanings of institutions, concepts, and values should, similarly, not be construed as a progressive building upon previous ones. Indeed, as we have already established, genealogy reveals that each meaning a thing takes on is nothing over and above it being arbitrarily "revaluated" by a new ruling class so as to reflect their own worldview and, therefore, validate their superiority: "anything having somehow come about is continually interpreted… to a new purpose by a power superior… every purpose… is just a sign that… [a new power] has impressed upon it its own idea" (GM 2:12). As such, genealogy reveals that major historical changes do not signal progress, but that a new ruling class has just put their own individual stamp on things.

Progress Despite Power: A Possible Objection to Nietzsche

We have established that Nietzsche's genealogy reveals things to have taken on different meanings in various civilizations, and that each meaning is a revaluation by means of different ruling classes in an ongoing game of power. However, what Nietzsche infers from this regarding history itself may not follow. By convincingly locating the nature and meanings of historically relevant institutions, values, and concepts by the interests and agendas they serve, Nietzsche contends that major historical changes should not be accounted for as progressively building on each other, but just a series of moves in that game of power. However, it seems that Nietzsche cannot sufficiently determine historical changes in meanings to not be progressing solely on the basis that they are always determined by and for some ruling class. That is, despite Nietzsche's contention that a power dynamic serves as the vehicle that moves history itself, we can still identify kernels of progress throughout history. Karl Marx, for example, agrees that history is the history

of power struggle (i.e., class conflict), yet Marx does not preclude the possibility of construing history as progressing insofar as the status of this power-relation itself has improved. As such, Nietzsche may have inferred to much; he has seemingly failed to entertain the possibility that despite history being moved by a power dynamic, that we can still identify, as I will show Hegel does, a thread progress in history. Ultimately, while Nietzsche offers no reason to infer that history is progressing, he does not seem to sufficiently show that history is not progressing.

Hegel's Conception of Spirit

Now I shall unpack Hegel's argument for why history is progressing in order to show that, despite Nietzschean anxieties concerning power dynamics, we can still identify a thread of progress throughout history. Hegel's philosophy of history is concerned with actual world history insofar as it is the "theatre" in which he thinks we are fundamentally witnessing the development of, what he calls, "Spirit". Here, Spirit's development tracks the development of human freedom: "the essence of Spirit is its freedom… Spirit is endowed with freedom… freedom is the only truth of spirit" (IPH 20). More specifically, Spirit's unfolding tracks the degree to which civilizations have become more conscious of and have, accordingly, realized their freedom where, here, "freedom" denotes the fact that, through and through, we are self-determining (IPH 21). That is, there is no immutable essence when it comes to Hegel's account of human freedom. Instead, humans and our freedoms are historical and changing; we are fundamentally the activity of producing ourselves and our world in history so as to bring about freedom. As such, Hegel thinks every civilization produces itself regardless of how aware they are of themselves as doing so, but Hegel's point is that *what* a civilization produces is determined by its Spirit because the extent to which a civilization is aware that it freely produces itself circumscribes the extent to which a

civilization can *actually* produce their world freely.

For Hegel, the state serves as the vehicle for Spirit to realize itself. As he puts it, "the State is the realization of freedom" (IPH 41). However, Hegel does not mean that a civilization's mode of government exhausts the institutions in which Spirit is realized, just that a civilization's mode of government serves as "the basis… for all [of a civilization's] spiritual activity" (IPH 52), i.e., Spirit is actualized in all a civilization's institutions, which makes sense given that Spirit circumscribes the bounds of what a civilization takes to be possible in the first instance. However, Hegel also that thinks civilizations mainly express their Spirit in their religion, art, and philosophy: "in religion… [Spirit] is represented, revered and enjoyed as God; in art… is depicted as an image and intuition; and in the philosophy is recognized and comprehended by thought" (IPH 55).

Hegel: History is Progressing

Now that we have established what Hegel means by 'Spirit', we are finally situated to understand the way in which he thinks that history is progressing. That is, *history is progressing* because, in the state in particular, we are witness to the progressive development of Spirit such that civilizations have become increasingly aware of their freedom, and by realizing Spirit in their institutions, humans have become freer. In other words, history is progressing insofar as the extent to which civilizations have actually freely produced their world has progressed. I should clarify, however, that Hegel does not conceive of history as progressing linearly; he fully admits that Spirit has progressed, stagnated, and even regressed to earlier stages. Instead of a wholly linear progress that all civilizations have contributed to, Spirit, and therefore history, is progressing only insofar as certain civilizations have structured their institutions to promote higher degrees of freedom (relative to what their predecessors were able to accommodate). Therefore, progress occurs only

when there has been a profound shift or deepening in our understanding of ourselves as free, self-determining agents.

If history is progressing in this way, we should be able to track Spirit across civilizations. Hegel offers such an analysis by tracking Spirit from Ancient Greece to Rome. In Greece, people were allowed to debate, but only if such debates did not challenge Greece's customs. However, once a new level of consciousness came about by way of Socrates (i.e., the Socratic method), people began to challenge their customs, which undermined Greek democracy. Ultimately, Greece's institutions were unable to accommodate what Socrates injected, which is why Hegel thinks Socrates was put to death and Greece fell (LPR). Indeed, a civilization cannot become more aware of their freedom while also maintaining institutions that do not promote such an awareness. Instead, institutions must progress to accommodate such higher levels of awareness or else they are at risk of falling like Greece. But just because Ancient Greece was unable to incorporate critical self-reflection into their institutions does not mean that Spirit did not develop, just that this more developed Spirit did not come to constitute the Spirit of Ancient Greece. Instead, Hegel traces this new level of freedom to constitute the Spirit of Ancient Rome, i.e., Ancient Rome's institutions were structured to promote critical self-reflection (LPR). As such, insofar as Rome's institutions were able to promote critical thinking while Greece's did not, Rome progressed farther than Greece, and insofar as Hegel is able to trace such progressive kernels throughout history, he thinks history is progressing.

Nietzsche: Calling Enlightenment Values into Question

Although Hegel's evaluation of progress in history seems more compelling than Nietzsche's (especially now that I have posited Hegel's account as an objection to Nietzsche), I

will now proceed to employ Nietzsche's genealogy to construct a rather devastating response to Hegel that aims to show not only that the Hegelian picture of progress poses no real threat to Nietzsche's stance, but it will also call Hegel's attempt to evaluate history into question altogether. As such, I will stop Hegel's philosophy of history from getting off the ground, which will allow me to conclude that Nietzsche's position is more convincing than Hegel's.

Nietzsche's consistent use of the terms "noble" and "slave" in describing the conflict I outlined earlier between the priests of Judea and the Romans suggests that Nietzsche thinks that Hegel, in his master-slave dialectic, fully admits that there has been a 'slave revolt' that is responsible for having injected, what have now become, enlightenment values of self consciousness and freedom into history. That is, insofar as Hegel thinks that history has been the history of Spirit's development where different civilizations have picked up the "torch", as it were, then he thinks that at some point in the past, these slaves picked up the torch and progressed past their masters, and that the development of Spirit ever since (or at least up till the enlightenment period) has been a history determined by the slaves and their values (i.e., a Nietzschean slave morality). As such, Nietzsche is suggesting that Hegel himself is the product of this slave history and so seems to have, on some level, dogmatically accepted and presupposed the terms on which the slaves understand things, i.e., Hegel seems to presuppose that freedom and self-consciousness serve as "the" universal stance by which history is correctly construed and evaluated.

Nietzsche's ultimate rebuttal to Hegel's philosophy of history is the following question: why should we accept Hegel's enlightenment values of freedom and self-consciousness as to be offering "the" universal standpoint from which history is correctly construed and evaluated? Indeed, Nietzsche is able to overcome my objection to him – that history can be construed as progressing despite Nietzsche's contention that a power dynamic serves as the vehicle that moves

history itself – because when Nietzsche asserts that history is not progressing, by no means is he denying that we can evaluate history as progressing if, for example, we were to judge it by Hegel's stance. Of course, there are various metrics by which we can evaluate history and its progress, but Nietzsche's point is that we cannot grant any such metric as "the" metric from which history is correctly construed in the first place? As such, a more accurate phrasing of Nietzsche's position is this: history is not progressing because *history cannot be evaluated as progressing.*

As Nietzsche's genealogy reveals by exposing the various ways in which Judea came to be injected with conflicting interpretations of what is "good" and "bad", there just is no metric contained in history itself that justifies it as "the" standard by which history can is correctly construed as "progressing" or "regressing". Rather, "good", "bad", "progressing", and "regressing", themselves are merely empty placeholders. That is, if such terms are to signify anything meaningful, that requires us to subject them to, and impress upon them, our own independent values such that we can come to evaluate, say, history as progressing or regressing. Furthermore, recall that by locating *who* injected certain interpretations of "good" and "bad" into the collective's imagination (and their motivations for doing so), genealogy reveals that throughout history, we have posited such interpretations and standards of evaluation because we have an invested interest in interpreting morality as such. In this same way, the only way we can grant positing Hegel's stance for evaluating the past over any other stance is because we have an invested interest in construing history according to that metric. As such, apart from our varying and independent interests in impressing certain values onto history, there just is no metric for evaluating history contained in history itself that justifies a normative claim such as progress. In this sense, history itself is not progressing because evaluating history as progressing is contingent on our differing and arbitrary interests in positing and granting this or that value as "the" stance

for evaluating history.

By calling any stance by which we might gauge progress in history, Nietzsche has undermined Hegel's entire attempt at evaluating history. Though, one might argue that insofar as Hegel's values are universal, so is his evaluation of history, but genealogy exposes precisely these enlightenment values as not universal, but just another arbitrary value system that functions to support a particular interest (e.g., to overthrow the Romans). Moreover, the possibility of Hegel's values serving some interest is not difficult to conceive of; Hegel's framework conveniently justifies imperialism. That is, Hegel thinks a more conscious and developed civilization can come to inject its awareness into a less aware civilization so as to "educate" them and improve the degree to which they produce themselves freely (IHP 87-89). Moreover, Hegel understands the British to have taken on this educative role (ibid.), and so it is at least plausible that Hegel's presupposed values simply function to legitimize the horrors of British colonization. Although genealogy does not sufficiently show that Hegelian values are not objective, it does strongly suggest it because we would have no reason to think that such values are inherently good given what we know about their genealogy.

Conclusion

I have offered a close reading of both Hegel and Nietzsche's accounts of whether history is progressing, and have made the case that although Hegel offers a convincing account for why history might be progressing (i.e., history as Spirit's unfolding), we can use Nietzsche's genealogy of morality to ultimately defend the Nietzschean position (that history is not progressing) by calling the values Hegel posits as "the" standard for evaluating history into question altogether and, thus, undermine Hegel's entire attempt to evaluate history as progressing. As such, I have

shown that the Hegelian picture of progress is not, as it is sometimes perceived, immune to severe criticism. However, readers may notice that this essay makes no attempt at challenging Nietzsche's genealogy on its own terms, and so this may be a possible avenue for responding to the Nietzschean position as I have presented it. Furthermore, this essay presents questions for future investigation. For example, should we be satisfied with Nietzsche's evaluation, or do we want to maintain that history can be seen as progressing? If so, this essay has made the case that we must seek an alternative strategy for doing so that does not fall prey to the Nietzschean critique.

Works Cited

Hegel, George Wilhelm Friedrich. *Lectures on the Philosophy of Religion.* Translated by R. F. Brown. Berkely: University of California Press, 1984.

Hegel, George Wilhelm Friedrich. *The Introduction to the Philosophy of History.* Translated by Leo Rauch. Indianapolis: Hackett Publishing Company, 1988.

Nietzsche, Friedrich. *On the Genealogy of Morality.* Translated by Carol Diethe, Cambridge: Cambridge University Press, 1994.

Understanding Mind Wandering: A Reply to Irving et al.

Zhiguo Huang — *Carnegie Mellon University*

Introduction

What is mind wandering? For years, philosophers and psychologists alike have attempted to characterize your experience on the couch reading a passage from a long and tedious novel. Your mind would sometimes drift away and think of lots of stuff: from planning what to have for lunch to criticizing the plot of the movie you watched last night. Folk psychology gives such experience its name - mind wandering. As a specific case of consciousness and mental action, mind-wandering has received burgeoning attention from researchers in the field of psychology and neuroscience.

Despite its simple outlook, however, mind-wandering episodes are not as homogeneous as they appear to be. Difficulties arise when we try to characterize its features, the necessary and sufficient conditions for mind-wandering states. This paper aims to answer this specific question: what counts as mind-wandering? How can we distill necessary and sufficient conditions for mind-wandering from current empirical research, thereby understanding this phenomenon on both personal and subpersonal levels? In section 2, I will first describe the phenomenon of mind wandering and previous attempts to define the phenomenon. Then in section 3, I will defend two features as necessary to the mind-wandering phenomenon, namely stimulus independence and absence of intention. In particular, this definition is a reply to Irving and Glasser (2020), who argue that mind-wandering can be stimulus-dependent and intentional. My definition provides an operational definition of mind-wandering and. I will also discuss how meta-awareness is involved

in mind-wandering, and why we should distinguish between possible types of mind wandering episodes. By cross-checking theoretical claims and empirical data, we may get a better understanding of mind-wandering as a specific form of consciousness.

Earlier definitions of mind-wandering

The phenomenon of mind-wandering, like almost all other psychological phenomena, has been characterized by William James. James used the following words to describe everyday experiences: "whilst part of what we perceive comes through our senses from the object before us, another part (and it may be the larger part) always comes out of our own head".[1] In other words, the "another part" here is independent of the external stimulus we receive through our perceptual system - which is our state during mind-wandering.

Earlier literature on mind-wandering defined it exactly as stimulus-independent or task unrelated thought. Mason et al. were the first to identify mind-wandering-correlated activity in the default mode network (DMN), an extended network in the brain that covers multiple regions, including the posterior cingulate cortex, angular gyrus, and large parts of the prefrontal cortex.[2] The activation of the default mode network can be seen in resting periods during wakefulness, especially in the absence of an external task. This finding has led to the characterization of "perceptual decoupling", or disengaging attention from perceptual stimuli, as a defining feature of mind-wandering.[3] Our train of thought is decoupled from perceptual input as our attention turns inward to the stimulus-independent thought, and it becomes harder to process or encode external stimuli. Task-independence is highly related to stimulus independence, and was also used as an

[1] William James, *The Principles of Psychology Volume II,* 103.
[2] Mason MF, Norton MI, Van Horn JD, Wegner DM, Grafton ST, Macrae CN. "Wandering minds: the default network and stimulus-independent thought," *Science*, 315 no. 5810 (2007):393-5.
[3] J. W. Schooler et al., "Meta-awareness, perceptual decoupling and the wandering mind," *Trends in Cognitive Sciences* 15, no. 7 (2011): 319-326.

operational definition of mind-wandering.[4] The stimulus/task-independence characterization was best articulated by Smallwood and Schooler, who define mind wandering as "a shift in the contents of thought away from an ongoing task and/or from events in the external environment to self-generated thoughts and feelings."[5]

Turning back to James again, the second part of his quote captures the spontaneity of mind-wandering: they seem to come out of our own head without external motivations. The spontaneous aspect of mind-wandering was proposed, as a complement to stimulus- and task independence, to account for the dynamic nature of mind-wandering. Smallwood and Schooler use "self-generated thoughts" to describe this aspect of mind-wandering. Intuitively, spontaneous thoughts, or actions in general, are the opposite of intentional thoughts or actions. When our minds start wandering, the thoughts that emerge seem non-intentional. In other words, we do not have a phenomenally conscious intention that causes the thoughts and experiences in mind-wandering. It is thus reasonable to say that mind-wandering is not intentional in general.

Meanwhile, in the more recent literature, the standard definition of mind-wandering has been challenged by various philosophers, which has been summarized in a review by Irving and Glasser.[6] They argue that their definition of "unguided attention" characterizes the dynamics of mind-wandering while the standard definition cannot. In the next section, I will argue that the standard definition can incorporate the dynamic aspect proposed by Irving and his colleagues, while not accepting the "unguided attention" proposal altogether.

[4] K. Christoff, "Undirected thought: Neural determinants and correlates," Brian Research 1428 (2012): 51-59.
[5] J. Smallwood and J. W. Schooler, "The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness." *Annual Review of Psychology* 66, no. 1 (2015): 488.
[6] C. Z. Irving and A. Glasser, "Mind-Wandering: A Philosophical Guide," *Philosophy Compass* 15, no. 1 (2020).

Reply to Irving and Glasser

Irving and Glasser review the definition of mind-wandering we have discussed above, and point out three difficulties facing the standard definition: (1) It fails to account for the dynamic nature of mind-wandering; (2) Mind-wandering can be related to a stimulus or a task; (3) Mind-wandering can be intentional. They argue that a better definition of mind-wandering is "unguided attention". First, I argue that mind-wandering must be unintentional, especially when we try to account for its dynamics. Then I examine the position they hold against stimulus and task-independence as necessary conditions for mind-wandering. Finally, I argue that a further demarcation of mind-wandering by the level of meta-awareness would facilitate empirical research. In this way, I show that the standard definition is still both phenomenologically accurate and empirically constructive.

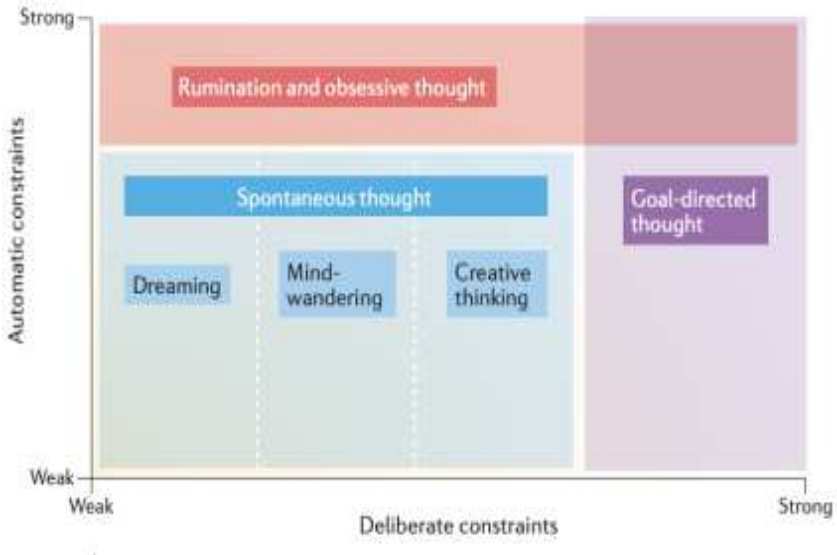
Spontaneity and the absence of intention

One concern against defining mind-wandering as strictly non-intentional is that some mind-wandering episodes can be "intentional" under some descriptions.[7] For example, when you are sitting at an hour-long lecture, at some point you might find listening to the lecture not sufficiently rewarding, and instead "intentionally" let your mind wander and think about other things. In the Davidsonian picture, an action is intentional if it is caused, in a non-deviant way, by some belief or desire.[8] The authors hold that under such description, the entire mind-wandering episode is intentional: you have a desire that "I want to let my mind wander because the lecture is so boring," which causes your mind to wander.

[7] Irving and Glasser, "Mind-Wandering: A Philosophical Guide."
[8] Donald Davidson, *Essays on Actions and Events (2nd ed.),* Oxford: Oxford University Press, 1980.

A short reply to this concern is that the presumed intention here is an intentional omission of control: we may intentionally not focus on the lecture.[9] However, since mind-wandering is phenomenally dynamic, we do not experience an intention that guides our thoughts through this automatic, generative process. Therefore, although in Irving's description, the whole mind-wandering episode seems intentional, in fact only its onset can be considered intentional in some way.

As in the discussion above, phenomenally, mind-wandering necessarily involves a dynamic process. The mind has to "meander" from one thought to another for the experience to be a mind-wandering episode, distinguishing it from rumination and obsessive thought.[10] Christoff et al. provide a dynamic framework that focuses on the spontaneous aspect of mind-wandering.[11] They argue that there are two types of constraints on the conscious stream of thought: the deliberate constraint, determined by the presence of an intention, and the automatic constraint, a family of mechanisms that operate outside of cognitive control. Spontaneous thought, including mind-wandering, arises when strong constraints are absent on both ends (see figure below).[12]



[9] S. Arango-Muñoz and J. P. Bermúdez, "Intentional Mind-Wandering as Intentional Omission: The Surrealist Method." *Synthese* 199 (2021):7727–7748.
[10] Irving and Glasser, "Mind-Wandering: A Philosophical Guide."
[11] K. Christoff et al., "Mind Wandering as Spontaneous Thought: A Dynamic Framework." Nature Reviews Neuroscience 17 (2016): 718–731.
[12] Adapted from Ibid., 719.

In Christoff et al.'s framework, rumination corresponds to high automatic constraint, while during mind-wandering, the automatic constraint is lower to allow for the "meandering."[13] If my reading is correct, Christoff and colleagues are essentially saying that mind wandering is unintentional and automatically dynamic. This characterization does not contradict the standard definition of mind-wandering as stimulus- and task-independent. Instead, it lead us to a definition of mind-wandering as an unintentional, dynamically developing train of thought that is stimulus-independent and task-unrelated.

The definition above is arguably a definition of mind-wandering that is less theoretically burdened than Irving and colleagues' definition of "unguided attention."[14] In the following section, I defend that the stimulus and task-independence definition as a successful one for mind-wandering.

Stimulus/task-independence

Irving and Glasser argued that mind-wandering can be stimulus-dependent using the following example:

> Consider Darnell, whose mind wanders on the bus to work. He smells delicious coffee, then imagines eating breakfast, then sees an insurance advertisement and remembers to check for quotes, then laughs at a remembered joke. While Darnell's mind wanders, he perceives stimuli in his environment: he smells coffee and sees an advertisement.[15]

Here, the authors seem to suggest that if the external stimuli play a causal role in somehow determining the content of mind-wandering episodes, then such episodes should be stimulus

[13] Christoff et al., "Mind Wandering as Spontaneous Thought: A Dynamic Framework."
[14] Z. C. Irving, "Mind-wandering is unguided attention: Accounting for the 'purposeful' wanderer," *Philosophical Studies* 173, no. 2 (2016): 547-571.
[15] Irving and Glasser, "Mind-Wandering: A Philosophical Guide," 2.

dependent. By this definition, Darnell's thoughts about eating breakfast and checking for quotes are stimulus-dependent. However, as their definition was laid out explicitly, it soon appears to be too inclusive to distinguish between stimuli that immediately evoke thoughts and stimuli that remotely lead to the same thought. This is less than satisfactory for an operational definition of stimulus-dependence.

I argue that to solve this problem, a stronger relation that involves an intention is needed between the external stimuli and the content of mind-wandering. For the characterization of stimulus-independence to be operational, we would rather define stimulus-dependence as the following:

> If agent A attends to external stimulus S and intentionally F's on S, then F is stimulus dependent on S.

By this definition, Darnell's wandering thoughts on coffee and the advertisement would not count as stimulus-dependent, because they, while being caused by the external stimuli, were not intentional actions or reactions performed on these stimuli. This definition should also be what previous researchers were thinking about when they define mind-wandering as stimulus/task unrelated thought, as they would also regard Darnell's case as stimulus-independent.

The review also points out that mind-wandering can be characterized as task-related in certain situations. More detailed arguments were put forth in previous papers by Irving and his colleagues. Irving and Thompson provide an example that our mind-wandering can be task-related: when a programmer's mind wanders to her code as she commutes home on a bus, her mind-wandering is related to her "task", which is coding. But again, Irving and Thompson are referring to a different relation between the task and mind-wandering content. They argue that

"tasks" in the real world are operationally defined as "whatever the person is currently doing."[16] If we accept this definition of task-relatedness, then Irving and Thompson are correct that it raises a problem: it is too permissive as a conceptual basis. Say I have a very remote task like "to retire at 40" or a very broad task like "to stay alive", then it seems that everything I do for now, by the authors' view, is related to these tasks. The problem here is analogous to a deviant causal chain: a task is remotely causing an action without properly causing the correct, immediate intention responsible for the action. Indeed, sometimes our minds do seem to wander to goals, but it does not necessitate that there are tasks that we actively take on. Instead of accepting this vague definition and admit the problems, however, we can modify the definition to support theoretical discussion. Specifically, when conceptualizing the phenomenon of mind-wandering, any phenomenological definition should account for its passive nature, which becomes the basis of my definition.

Similar to stimulus-dependence, I propose a definition of task-relatedness that better captures the phenomenology of mind-wandering as unintentional:

> If for a task T, an intention S to achieve T causes agent A to perform an intentional action F, then F is task-related to T.

Now under our definition, we will not be forced to say that everything we do now is related to the task "to survive," because they are not caused by an intention that literally says "to survive." The programmer's mind-wandering to her codes would also be task-unrelated because it isn't caused by an intention to program. This definition is arguably more useful for future research on mind-wandering because it properly defines the widely accepted intuitive view that mind wandering is task-unrelated.

---

[16] Z. C. Irving, and E. Thompson, "The Philosophy of Mind-Wandering" in *The Oxford Handbook of Spontaneous Thought: Mind-wandering Creativity and Dreaming.* Edited by K. Fox and K. Christoff. (Oxford: Oxford University Press, 2018), 89.

We have now established that under our improved definition of stimulus-dependence and task-relatedness, mind-wandering is necessarily stimulus- and task-independent, contrary to the conception of Irving and his colleagues. It should be noted that our definition of stimulus and task-dependence focuses on the role that attention/intention, or rather the lack thereof, plays in mind-wandering. This is indeed an important aspect, and we will explore it further in the next section.

Two Types of Mind-Wandering

For a less standard definition of mind-wandering, Metzinger (2013, 2015) hold that in a mind-wandering episode, the agent lacks veto control, which requires meta-awareness. Irving and Glasser rejected Metzinger's view because meta-awareness can be present in mind wandering episodes.[17] Smallwood and colleagues suggested that mind-wandering with or without meta-awareness are two different mental activities. They used the following descriptions to provide instructions to their participants:

> **Tuning Out**: Sometimes when your mind wanders, you are aware that your mind has drifted, but for whatever reason you still continue to read. This is what we refer to as "tuning out"—i.e., when your mind wanders and you know it all along.
>
> **Zoning Out**: Other times when your mind wanders, you don't realize that your thoughts have drifted away from the text until you catch yourself. This is what we refer to as "zoning out"—i.e., when your mind wanders, but you don't realize this until you catch it.[18]

The behavioral evidence was confirmed by neuralimaging studies. fMRI evidence suggests that the brain network activated during mind-wandering episodes with meta-awareness is different

[17] Irving and Glasser, "Mind-Wandering: A Philosophical Guide."
[18] J. Smallwood, M. McSpadden, and J. W. Schooler, "The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders." *Psychonomic Bulletin & Review* 14, no. 3 (2007): 533.

from those without meta-awareness.[19] In the non-aware cases, high levels of activation can be seen in major components of the default mode network, whereas in the aware cases, the activation levels are much lower.

Categorizing mind-wandering into aware and non-aware types have two implications for empirical studies. First and foremost, the difference in phenomenal character suggests that future research should further specify the functional roles of the default mode network and identify more accurate neural correlates. But more profoundly, a comparison between aware and non aware mind-wandering episodes can respond to a concern in current introspective studies, which has been referred to as the paradox of introspective report.[20] Most of the existing studies with experience sampling ask subjects to introspect and report their experience. However, the accuracy of these subjective reports is limited because they disrupt the natural evolution of experience, especially in a dynamic process like mind-wandering. These problems highlight a paradox in studying mind-wandering: metacognitive access is (almost) necessary for studying conscious mental states, but the access itself alters the experience and leads to problems in the data. If we can operationally distinguish between aware and non-aware mind-wandering episodes in experimental settings, we can potentially investigate the effect of meta-awareness on introspective reports.

Conclusion

In this paper, I approach the conscious mental states of mind-wandering from both philosophical and empirical perspectives. On the philosophical side, mind-wandering can be

[19] K. Christoff et al., "Experience sampling during fMRI reveals default network and executive system contributions to mind wandering." *Proceedings of the National Academy of Sciences* 106, no. 21 (2009): 8719-8724.
[20] M. Konishi and J. Smallwood, "Shadowing the wandering mind: How understanding the mind-wandering state can inform our appreciation of conscious experience," *Wiley Interdisciplinary Reviews: Cognitive Science* 7, no. 4 (2016): 233–246.

defined as a non-intentional, dynamic succession of stimulus-independent, task-unrelated thoughts. I defend stimulus-independence and task-unrelatedness as necessary features using our improved definition. This characterization is then combined with an account of the dynamic, automatic aspect of mind-wandering from Christoff and colleagues to form the definition I propose, which is arguably sufficiently simple and illuminating to direct future research in psychology and neuroscience. On the empirical side, my definition of mind-wandering yields a more practical paradigm that fits in ongoing studies in psychology and neuroscience. It corresponds to the proposal of the default mode network as a neural correlate of stimulus independent, task-unrelated thoughts, and suggests improvement in current experimental paradigms to account for the paradox of introspection and to specify the relation of meta awareness to different neural activation patterns.

This paper should shed some light on the future directions in studying mind-wandering experiences. Using a simpler definition allows neuroscientists to design more specific experiments, and the potential relation between meta-awareness and mind-wandering asks for better experimental paradigms. The relation between intention or agentive control and mind-wandering might also be a fruitful topic to explore in further research as implied in our discussion.

Works Cited

Arango-Muñoz, S. and Bermúdez, J. P. "Intentional Mind-Wandering as Intentional Omission: The Surrealist Method." *Synthese* 199 (2021):7727–7748. https://doi.org/10.1007/s11229-021-03135-2.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., and Schooler, J. W. "Experience sampling during fMRI reveals default network and executive system contributions to mind wandering." *Proceedings of the National Academy of Sciences* 106, no. 21 (2009): 8719-8724. https://doi.org/10.1073/pnas.0900234106.

Christoff, K. "Undirected thought: Neural determinants and correlates" *Brian Research* 1428 (2012): 51-59. https://doi.org/10.1016/j.brainres.2011.09.060.

Christoff, K, Irving, Z. C., Fox, K. C. R., Spreng, R. N., and Andrews-Hanna, J. R. "Mind Wandering as Spontaneous Thought: A Dynamic Framework." *Nature Reviews Neuroscience* 17 (2016): 718–731. https://doi.org/10.1038/nrn.2016.113.

Davidson, D. *Essays on Actions and Events* (2nd ed.), Oxford: Oxford University Press, 1980.

Irving, Z. C. "Mind-wandering is unguided attention: Accounting for the 'purposeful' wanderer." *Philosophical Studies* 173, no. 2 (2016): 547-571. https://doi.org/10.1007/s11098-015- 0506-1.

Irving, Z. C., and Glasser, A. "Mind-Wandering: A Philosophical Guide." *Philosophy Compass* 15, no. 1 (2020). https://doi.org/10.1111/phc3.12644.

Irving, Z. C., and Thompson, E. "The Philosophy of Mind-Wandering" in *The Oxford Handbook of Spontaneous Thought: Mind-wandering Creativity and Dreaming*. Edited by K. Fox and K. Christoff. Oxford: Oxford University Press, 2018.

Konishi, M., and Smallwood, J. "Shadowing the wandering mind: How understanding the mind-wandering state can inform our appreciation of conscious experience." *Wiley Interdisciplinary Reviews: Cognitive Science* 7, no. 4 (2016): 233–246. https://doi.org/10.1002/wcs.1392.

Mason MF, Norton MI, Van Horn JD, Wegner DM, Grafton ST, Macrae CN. "Wandering minds: the default network and stimulus-independent thought." Science. 2007 Jan 19;315(5810):393-5. doi: 10.1126/science.1131295. PMID: 17234951; PMCID: PMC1821121.

Mittner, M., Hawkins, G. E., Boekel, W., and Forstmann, B. U. "A Neural Model of Mind Wandering*." Trends in Cognitive Sciences* 20, no.8 (2016), 570–578. https://doi.org/10.1016/j.tics.2016.06.004.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., and Sayette, M.

A. “Meta-awareness, perceptual decoupling and the wandering mind.” *Trends in Cognitive Sciences* 15, no. 7 (2011): 319-326. https://doi.org/10.1016/j.tics.2011.05.006.

Smallwood, J., McSpadden, M., and Schooler, J. W. “The lights are on but no one’s home: Meta-awareness and the decoupling of attention when the mind wanders.” *Psychonomic Bulletin & Review* 14 no. 3 (2007): 527–533. https://doi.org/10.3758/BF03194102.

Smallwood, J., and Schooler, J. W. “The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness.” *Annual Review of Psychology* 66, no. 1 (2015): 487–518. https://doi.org/10.1146/annurev-psych-010814-015331

James, William. *The Principles of Psychology.* New York: Henry Holt & Co, 1918.

Grasping Virtue

Hassan Saleemi — *University of Illinois at Chicago*

Introduction

In *Natural Goodness*, Philippa Foot argues for the compatibility of the objectivity and motivating influence of moral judgments. She argues for the first by "likening the basis of moral evaluation to that of the evaluation of behavior in other animals."[1] Just as there is something wrong with the pigeon that cannot fly, there is something wrong with the liar. The liar's moral defect is identical with this natural defect, a defect in their rational will. These judgments are practically motivating because they are judgments about what a human being ought to do rationally.[2] In saying the liar is defective, one may be saying that the liar ought to tell the truth on pain of practical irrationality. Skepticism has been raised about both parts of this view, but my primary concern is skepticism about the objective truth of moral judgments. Call it *objectivity skepticism*. The objectivity skeptic's claim is that it is implausible that 'good' and 'bad' could be univocal between the ethical and ethological contexts because human life is radically different from animal life—so ethical goodness cannot be natural goodness of the human will. I believe objectivity skepticism rests on a particular understanding of virtue that we should reject in which virtue is nothing more than the recognition of a moral requirement (this will be made more precise). In providing my alternative view about how we should understand virtue, I defuse both objectivity skepticism and skepticism about the ability of moral judgments to motivate (call it *practicality skepticism*). The

[1] Philippa Foot, *Natural Goodness*, (Oxford: Oxford University Press, 2001), 16.
[2] Ibid., 9.

essay proceeds in three stages. First, I develop the metaphysical background. Second, I explain objections to Foot's view and answer them by modifying Foot's picture of virtue. Third, I answer objections to my modifications.

Background

Foot's view is predicated on Michael Thompson's analysis of the concept of life. According to Thompson, our knowledge of an organism as a living thing is knowledge of that organism in terms of a life-form of which that organism is an instance. Our belief that there are human beings rests on a conception of human beings as individuals rather than as colonies of cells. And this conception of human beings as individuals draws attention to the question: individual *whats*? The answer is: instances of a universal form of life.[3]

Foot believes this picture can be exploited to meta-ethical ends. The concept of a form of life is a metaphysical one that makes possible certain judgments.[4] Of interest to Foot are the judgments Thompson calls natural-historical judgments or Aristotelian categoricals.[5] Here are some examples: human beings learn language in an early critical period, human beings have four-chambered hearts, human beings have wisdom teeth. These categoricals have interesting properties. Most obvious in the categorical regarding language is the tense; the categoricals possess a unique timelessness in virtue of being true of an organism's natural history or life cycle. Most obvious in the categoricals about hearts and teeth is the strange kind of generality; the categoricals

[3] Michael Thompson, *Life and Action*, (Cambridge: Harvard University Press, 2008), 56-62; Foo, *Natural Goodness*, 27-8.
[4] Michael Thompson, "Apprehending Human Form." *Royal Institute of Philosophy* 54 (2004): 63.
[5] Thompson says that Aristotelian categoricals are sentences which express the propositions he calls natural-historical judgments (Thompson, *Life and Action*, 64 - 65). Foot uses the latter term instead of the former and so I will simply follow her use since it is her work I am concerned with.

have this non-Fregean generality in virtue of being characteristically (rather than universally) true of a form of life. It is true that human beings have wisdom teeth even if everyone gets theirs removed.[6]

But Foot doesn't want us to worry about wisdom teeth. She only cares about categoricals that mark out a teleology. Unlike wisdom teeth, which are vestigial, language and cardiac structure matter in that, if one cannot communicate or one has a congenital heart defect, one is defective *qua* the human form of life. And just as there is natural defect, there is natural goodness: athletes in excellent cardiovascular health are excellent in this respect. The trait predicated of the form of life in an Aristotelian categorical has the relevant sort of teleology iff it contributes to the organism's being a good instance of that form of life. For the vast majority of organisms, this is cashed out as being well-placed to grow, develop, and reproduce.[7] Borrowing terminology from Elizabeth Anscombe, Foot calls the necessity that organisms have the capabilities in order to be a good instance of that organism an Aristotelian necessity. So we can put the point in this way: Foot is interested in Aristotelian categoricals that represent an Aristotelian necessity. I use "Aristotelian categorical" to refer to this subset.[8]

Problems

The thesis in question is Foot's thesis that ethical goodness is the natural goodness of the human will.[9] Recall the pigeon that cannot fly; recognizing the Aristotelian necessity of flight for pigeons, we truthfully claim that the pigeon is defective—bad as a pigeon. Similarly, we say that

[6] Thompson, *Life and Action*, 19-21.
[7] Foot, *Natural Goodness*, 30-3.
[8] Ibid., 46.
[9] Ibid., 16, 39.

the liar is bad as a human being—a bad human being. And what's more, we only need the light of objective facts to see this.

Objectivity skepticism has its roots in the observation that human life is radically different from animal and plant life. Whereas for most organisms, being a good instance of one's form of life is cashed out in terms of fitness, a human life limited to this biological notion is not a life anyone would want to live.[10] Foot herself recognizes this:

> What conceptually determines goodness in a feature or operation is the relation, for the species, of that feature or operation to survival and reproduction, because it is in that that good lies in the botanical and zoological worlds … at that point questions of "How?" And "Why" and "What for" come to an end. But clearly this is not true when we come to human beings.[11]

All this points to the conclusion that ethical goodness cannot be natural goodness; we cannot account for what is good in human life in naturalistic terms.[12] Foot attempts to deal with these difficulties by giving a negative account: a human life is naturally good insofar as it lacks some natural defect.[13] But this is problematic. First, it fails to capture what is positively distinctive about human life, a significant handicap on any attempt to ground the virtues (kindness is not simply a lack of cruelty). Second, it seems to concede that natural goodness is an inadequate concept. If we cannot extend natural goodness to what is distinctive about human life, we are admitting it is not enough.

Matthias Haase notes that there is an easy solution for Foot. We need only recognize that natural goodness and natural defect are formal concepts substantiated differently in different forms of life. We need not say that the teleology specified in judgments about human beings goes beyond "the cycle of self-maintenance and reproduction"; we only need to say that what is involved in this

[10] But even for animals these are not enough: play, social interaction, etc. are also part of some animals' form of life.
[11] Foot, *Natural Goodness,* 42.
[12] Ibid., 43.
[13] Ibid., 43-4.

formally specified cycle takes on a distinctive substantive shape.[14]

In this essay, I defend the univocity of natural goodness. But Haase thinks that we cannot say that natural goodness is a univocal concept because of the consequences he describes here:

> The problem with [the view associated with the univocity of natural goodness is that] if the general teleological framework for thinking about living things is the *same* no matter which vital powers come into view, then it would seem that the "method" of establishing what is good for a human being must indeed be the same as [the method for the other animals]. The latter is clearly an empirical investigation that rests on observation … and if that is so, then the power of practical reason will always put me in the position to "step back" and ask: "Why should I do what humans do?" … Practical reflection, therefore, cannot find its ground in the *given* necessities of life.[15]

Haase articulates two concerns here. First, Haase is echoing John McDowell's concern that Foot turns ethics into an empirical science with a grounding "from the outside."[16] It might make one think Foot is arguing for a "vulgar evolutionary ethics … [turning] ethics into a sub-discipline of biology."[17] If moral knowledge is known through empirical investigation, it is "known in the wrong way."[18] Both Haase and Thompson cite McDowell in making their points. But I don't see why Foot cannot, in principle, make ethics into such a science so long as she is able to show, as I will argue, that the findings of this science are *what it is rational to do*.

The second concern is that there is no binding reason to act ethically if such knowledge is empirical. We can always step back and ask why human beings should manifest the behavior found in Aristotelian categoricals. It is only if we understand moral knowledge from the "perspective of thought and choice", that we are bound. This problem is a form of moral skepticism distinct from

[14] Matthias Haase, "Practically Self-Conscious Life." *Philippa Foot on Goodness and Virtue*, Ed. John Hacker-Wright. (London: Palgrave-Macmillan. 2018.), 102-3.
[15] Ibid., 104-5.
[16] John McDowell, "Two Sorts of Naturalism," *Virtues and Reasons: Philippa Foot and Moral Theory*, Ed. Rosalind Hursthouse, (Oxford: Clarendon Press. 1995),
[17] Michael Thompson, "Apprehending Human Form," 62-3. Cf. Thompson, "What is it to Wrong Someone," 377.
[18] Haase, "Practically Self-Conscious Life," 104.

the two I introduced already. Not only must moral judgments be motivating, but being so motivated must be justified—call it *justification skepticism.* Haase's idea is that if objectivity skepticism is dealt with in such a way that maintains the univocity of natural goodness, then Foot is open to justification skepticism because of the empirical method it requires. Let me explain.

Justification skepticism is challenging in light of a scientific understanding of reality that seems to leave no room for the rationality of ethical considerations. McDowell's move is to say that we should "stop supposing the rationality of virtue needs a foundation outside the formed evaluative outlook of a virtuous person."[19] Haase's move is to argue that natural goodness is a distinct concept in the realm of the human. As we ascend the *scala naturae*, there is "a transformation of the shape that the cycle self-maintenance and reproduction takes" in not just the substantive nature of natural goodness, but its formal nature as well.[20] This allows him to follow Thompson in saying that the method for grasping what natural goodness consists in for human beings is through an *a priori* understanding of one's own practical activities rather than an empirical one.[21]

What is Foot's move? There is one obvious way to understand justification skepticism and one non-obvious way. The obvious way is to understand the question as asking how Foot grounds the rationality of acting virtuously on the *de dicto* reading of 'acting virtuously.' The question is: what makes virtuous action rational? I have mentioned at the beginning of this essay that Foot's way of coping with practicality skepticism is to argue that moral judgments are judgments about what we rationally ought to do. If she is successful, then it seems she has answers for both practicality skepticism and justification skepticism. But there is a non-obvious way that I think

[19] McDowell, "Two Sorts of Naturalism."
[20] Ibid., 112-9.
[21] Haase, "Practically Self-Conscious Life," 120-3.

does pose a problem: to understand the question as asking how Foot grounds the rationality of acting virtuously read *de re*. The question is then: what makes these actions, which are virtuous, rational? To explain the problem, I must introduce what Haase calls the Apprehension Requirement.

The Apprehension Requirement is the requirement that a meta-ethical theory must account for the truth of this claim: "when 'good' features in 'good human action', it implies that the subject acts on an understanding of what is good to do."[22] Ethical goodness is practically self-conscious in that in it, the subject is aware of their grounds for acting ethically. The idea is that a meta-ethical theory must be able to explain the connection between an agent's ethically good intentions and the fact that what they do is ethically good. If ethical goodness is the natural goodness of the human will, then the connection is naturally articulated like so: when someone keeps their promises, it is because they recognize that keeping one's promises is part of what makes a human being good *as* a human being—part of what it means to have a rational will. We must avoid over-intellectualization here—all this requires is that, for example, someone refuses to steal because they have the thought "I can't take this; it's hers."[23] But still, it is the recognition of the goodness or badness of the action in question; if the person is made to explain why the thought "it's hers" has the force it does, reference to natural goodness is what would make their explanation correct. The point of mentioning the Apprehension Requirement is this: if natural goodness is univocal, then according to Haase this requires that one recognizes the goodness of promise-keeping through empirical investigation. *That* is the way the connection we just drew is formed in the individual. But that seems implausible. For one, empirical investigation will never reveal to us the ethical status of some actions: not every action which is good is such that good hangs on it. And second,

[22] Haase, "Practically Self-Conscious Life," 96.
[23] Ibid.

no one learns the full extent of right and wrong through acts of ostention. It seems so implausible that we might want to reject that 'naturally good' is univocal and claim that we can grasp what is naturally good for human beings in another way, as Haase does. This, I think, is the real motivation for the position which Haase takes. This is why Foot is open to justification skepticism.

Thus we have justification skepticism *de re*: it seems obscure how we are to understand the truth of the claim that we have a reason to act in this particular virtuous way such that Foot could meet the Apprehension Requirement—for that requires an explanation for how one *could* come to grasp the goodness of acting in that particular way for every action, even if that is not how one does grasp it. As things stand, it seems that either Foot rejects the univocity of natural goodness so she can meet the Apprehension Requirement in the way Haase argues she should, or she accepts the univocity of goodness and fails to meet the Apprehension Requirement.

Foot's Picture of Virtue

Objectivity skepticism is predicated on justification skepticism and justification skepticism is skepticism about whether Foot can meet the Apprehension Requirement given the univocity of natural goodness. We are not at the bottom floor yet. I think that the dilemma seems live because we impute Foot with accepting a particular picture of how virtue figures in human life. But we are better off without this. The connection between Foot's difficulty in meeting the Apprehension Requirement and this picture of virtue is best articulated by Anselm Müller:

> Some oughts express … [a] requirement to implement a certain motivational pattern, ideally consolidated in a virtue of character. Such a pattern typically connects a motivating reason with a kind of action required by the presence of that reason … [but] it is by no means obvious that 'the good person' does have … a reason in favor of being motivated as virtue requires … where does such a reason come from? Natural Normativity?[24]

[24] Anselm Muller, "'Why Should I?' Can Foot Convince the Skeptic?" *Philippa Foot on Goodness and Virtue*, Ed. John Hacker-Wright, (London: Palgrave-Macmillan, 2018), 155, 158.

The question Muller asks at the end of this passage is another way of formulating the question of how Foot meets the Apprehension Requirement: what about a person's intentions could make their actions good? What reason could they be acting on? Muller is unclear on how this is possible[25] because he conflates "a reason in favor being motivated as virtue requires" with a reason to implement a motivational pattern (he calls such a reason a motivational requirement)—this is what it seems he must do if he accepts the claim that a virtue is something which consolidates these motivational patterns. A motivational requirement is a reason to accept the motivational pattern characterized by the conditional: "if state of affairs S obtains, treat S as a reason to A." Muller's example is of the motivational requirement not to lie: "if one knows not-*p*, treat this knowledge as a reason to not say *p*." Since Muller conflates a reason to be motivated as virtue requires with a reason to accept a particular motivational requirement like this, he might naturally turn to the concept of particular patterns of natural normativity (behavior represented in Aristotelian categoricals) in formulating his rhetorical answer to the question of how Foot meets the Apprehension Requirement. And those patterns are discovered through empirical investigation. Haase shows that he shares this way of understanding Foot when he writes "what being a just person requires is the recognition of … patterns of rationalization" involved in Aristotelian categoricals such as 'human beings recognize rights.'[26]

The problematic picture is this one: a virtue consists in the implementation of particular motivational requirements consolidated in that virtue. Foot seems to believe this sometimes. She writes of particular courses of action as being required by their Aristotelian necessity such as the

[25] Muller is asking this question rhetorically. He, like McDowell, does not actually think that we can or should try to justify acting in accordance with virtue to someone who is not already virtuous. Foot does care (Foot, *Natural Goodness,* 53, 64 - 65) and so do I. So I am not criticizing Muller here because he is unclear on how Foot could do this—he doesn't think anyone can—I am criticizing his interpretation of Foot that makes it make sense to ask this rhetorical question.
[26] Haase, "Practically Self-Conscious Life," 96.

keeping of promises and the prohibition of murder and theft.[27] She also claims that "the distinguishing characteristic of the just [is] that for them certain considerations count as reasons for actions, and as reasons of a given weight"[28] —that they recognize certain motivational requirements. If we understand Foot to hold that this is what a virtue like justice is, it seems obvious that she should meet the Apprehension Requirement by holding that the way the connection is created between good intentions and good actions is through recognition of particular motivational requirements. Then the univocity of natural goodness and the empirical method which it implies we must use does indeed pose a problem. For example, Müller says of the categorical "human beings let harmless animals live" that "it seems not to be excluded by an Aristotelian necessity."[29] And if recognition of Aristotelian necessities is the only way someone might let harmless animals live and be acting rightly, then the univocity of natural goodness is untenable. Foot's account must fail.

Christine Korsgaard's criticisms of Foot help explain how all this has come to pass. She argues that if virtue just is being practically rational in certain circumstances—if it just is recognizing particular motivational requirements in action—then the notion of virtue seems to do no real philosophical work.[30] To say someone had a virtue would then just be to name a particular fact about the way they act: that they recognize those motivational requirements. And if that is the case, why not just talk about the motivational requirements themselves? This, in fact, is what Haase and Muller seem to do. Luckily, we can get by without doing this.

[27] Foot, *Natural Goodness*, 45, 114.
[28] Ibid., 12.
[29] Muller, "Why Should I?" 164.
[30] Christine Korsgaard, "Constitutivism and the Virtues," *Philosophical Explorations* 22 (2019): 20.

Re-Interpreting Foot

My solution involves providing an alternative understanding of virtue. Despite the significant textual support for the interpretation Haase, Muller, and Korsgaard seem to hold, there are also indications of this other picture in Foot's work. Nevertheless I should be clear that I am not claiming that this is Foot's actual view.

In developing objectivity skepticism, I started from Foot's thesis that ethical goodness is the natural goodness of the will. We saw rather quickly, however, Foot's thesis about the justification of acting well come into view. This is because the picture of virtue Foot provides connects these two central theses. So let me consider them together:

(1) Ethical goodness and badness are natural goodness and defect respectively in the human will.[31]
(2) "Goodness sets a necessary condition of practical rationality and is at least a part-determinant of the thing itself."[32]

I read Foot as believing that (2) features in an explanation of the truth of (1) and as believing that (1) features in an explication of the meaning of (2)—that is, what it means in the human case that goodness determines the nature of practical rationality. Because (1) requires the truth of (2), I will start there.

Foot's defense of (2) is an analysis of virtue. But before I turn to her argument, I should say that, though I use the example of promise-keeping throughout, this is only a heuristic. The thesis is a formal one about practical rationality wherever it is found—the goodness of the will of that form of life will determine what it looks like. (1) will be required to make sense of the human case.

In order to understand the way goodness might determine what practical rationality might

[31] Foot, *Natural Goodness*, 16, 39.
[32] Ibid., 10-1, 63.

look like, we need to understand goodness—goodness of character in general, which involves such things as prudence. Foot describes the virtues which determine the nature of practical rationality in terms of "(a) the recognition of the particular considerations as reasons for acting and (b) the relevant action" performed in the light of those reasons.[33] If we stop here, we get the reading of virtue as the recognition of motivational requirements, which would lead to the various problems I have discussed. For if having a virtue just is being practically rational in certain circumstances, there is nothing left of virtue to determine the nature of practical rationality. This is how Korsgaard's criticism gets off the ground. So we are fortunate that Foot says more about virtue than this. She claims that "the underlying attitudes and desires [involved are] an essential part of virtue."[34] In writing about promise-keeping, Foot turns to a figure in Kropotkin's *Memoirs of a Revolutionist* and says this of him:

> Why *should* he have kept his promise? How do *good* and *bad* come in here? … Promises belong to the area of trust and respect for others … disrespect and untrustworthiness are bad human dispositions. It matters in a human community that people can trust each other, and matters even more that at some basic level humans should have mutual respect. It matters, not just what people do, but what they are.[35]

This account seems to imply that the goodness of keeping one's promises is rooted in the goodness of the dispositions involved. But as I pointed out briefly in the previous section, Foot writes that promise-keeping is something human beings ought to do because good hangs on our ability to bind each other's wills.[36] How are we to reconcile these two accounts?

---

[33] Foot, *Natural Goodness*, 13.
[34] Ibid., 113.
[35] Ibid., 48. In a footnote in 'Rationality and Goodness', Foot denies being a virtue ethics theorist if holding such a view requires that one believe that dispositions are the source of goodness and badness. This contradicts how I am choosing to read her and what this quotation seems to say. This quotation seems to imply that the source of goodness and badness is human dispositions. I don't know how to make sense of this unclarity in Foot's view other than to downplay the significance of this quotation if our purpose is historical accuracy.
[36] Ibid., 45.

Foot writes that the derivation of the relevant categorical is rooted in the good that hangs on some course of action. Knowing that good hangs on promise-keeping is enough to argue:

> Human beings keep their promises.
> This human being does not keep their promises.
> Therefore, this human being is defective.

But there are two problems. First, this does not connect promise-keeping with practical rationality—it does not show that we *ought* to keep our promises and thus leaves Foot open to justification skepticism.[37] Second, in order to understand "how *good* and *bad* come in here" we need something more than this theoretical conclusion. My interpretation of Foot does not connect promise-keeping with rationality directly, but takes a detour through virtue. Drawing this connection requires the validity of the following form of inference using Aristotelian categoricals:

> The life-form S is/does/has F.
> Being/Doing/Having G is a necessary condition of being/doing/having F.
> Therefore, the life-form S is/does/has G.

This is plausible. And so we can write:

> Human beings keep their promises.
> Being just is a necessary condition of reliably keeping one's promises.
> Therefore, human beings are just.[38]

I will devote the next section to the soundness of this inference. The result will be that some alterations must be made. For now, it is enough to recognize that, if sound, we've arrived at the Aristotelian necessity of having a *virtue*. Foot mentions that good hangs on specific virtues, but never shows how these Aristotelian necessities are derived.[39] Something like the reasoning I have

---

[37] The poverty of the logic of Aristotelian categoricals is stressed especially by McDowell in 'Two Sorts of Naturalism.'

[38] Two things. First, I recognize that the second premise here is not obviously true and will come back to it later. Second, (1) has, as all categoricals do, non-Fregean generality. It is hard to capture this generality in a proposition which is not an Aristotelian categorical. We can do our best to capture this in English by introducing the adverb 'reliably' in (2). Later I replace this with 'characteristically.'

[39] Foot, *Natural Goodness*, 44-5.

just presented seems to be required, for one can only find out that good hangs on the virtues for our form of life if good hangs on the empirically identifiable manifestations of those virtues. The Aristotelian necessity of having a virtue will be understood in accordance with the previous quotations from Foot as the Aristotelian necessity of having a certain disposition that (at least) involves various mental states (attitudes, beliefs, desires, etc). And because the concept of a virtue is ethically loaded from the get-go, it is clear to see why this man *should* have kept his promise. But it is not yet clear that he would be irrational if he did not.

In distinguishing the possession of a virtue from being practically rational, we have created for ourselves the need to take a second step: showing how virtues *determine* the nature of practical rationality. Specifically, virtue must be connected with practical rationality in such a way that we can say that the person who acts immorally acts irrationally. So I say: those courses of action are irrational which the idealized virtuous agent could not have chosen, all things considered. The justification for saying this is that otherwise practical rationality could not be important. Following Warren Quinn, Foot argues that we could not be right in caring about being practically rational if an instrumental theory of practical rationality were true. The thought is that, if we could accomplish any evil and still be practically rational, rationality in itself could not be a virtue—at the very least, we would be rather ambivalent about it.[40] This observation opens up the door for this way of arguing: rather than starting off with a theory of practical reasoning and trying to see how ethics can be wrought from it, we should start off by "seeing goodness as setting a necessary condition of practical rationality and therefore at least a part-determinant of the thing itself."

Now we can turn to thesis (1). Foot's way of arguing for this thesis is by showing that the derivation of the Aristotelian necessity of promise-keeping mirrors the derivation of the

[40] Foot, *Natural Goodness*., 10, 62-3.

Aristotelian necessity of certain acts and features of plants and animals—she attempts to show that ethical goodness is a kind of natural goodness.[41] But this way of arguing has to assume already that keeping one's promises is something which is ethically good; otherwise, how could it show that the derivation of ethical goodness mirrors that of natural goodness? I am not going to provide a new reading of Foot's argument for this thesis, but we do not need to assume that any particular act is ethically good in order to argue for (1). And we do not need to do this because we already have (2) in place (Foot reverses the order of argumentation in *Natural Goodness*). Given (2), it is clear that practically rational action is virtuous action. And if that is the case, the following argument can be made:

> The natural goodness of the human will consists in practically rational action.
> Practically rational action is virtuous action (from (2)).

Therefore, the natural goodness of the human will consists in virtuous action. In other words, the natural goodness of the human will consists in its ethical goodness. This is (1). In addition to mentioning that (2) features in an explanation of the truth of (1), I mentioned that (1) features in an explication of (2) in the human case. All I mean by this is that, since (2) only tells us about practical rationality in the abstract, we need a more concrete notion of ethical goodness in order to figure out what practical rationality looks like in the human case. And since the conclusion of the proof just given defines virtuous action in terms of the natural goodness of the human will, that is what we now have. Given how closely linked (1) and (2) are, it is no surprise that a misunderstanding regarding (2) could lead to skepticism regarding (1). Before I move on to how this account will solve the skeptical worries that were previously raised, I will now turn back to what I have just glossed over: the inference to the Aristotelian necessity of a virtuous disposition.

[41] Foot, *Natural Goodness*, 46.

Two Objection

The understanding of Foot presented relies on the idea that we can infer the Aristotelian necessity of a virtue from the Aristotelian necessity of some action. Does this work? Here is the inference that we are considering:

(1) Human beings keep their promises.
(2) Being just is necessary condition of reliability keeping one's promises.
(3) Therefore, human beings are just.

The problem is that (2) is not obviously true; it will likely strike many as obviously false. Being just, it seems, is merely a sufficient condition for reliably keeping one's promises. Another sufficient condition for reliably keeping one's promises is being fearful of social sanctions. More plausibly, the objector might define a notion like justice* which is narrower in certain respects than justice. Justice*, they might say, ensures the reliable production of this action, but not of adjudicating fairly among one's equals or distributing goods equitably, or any of the other sorts of things we associate with justice. For the reliable production of these latter courses of action, we need justice** and justice***. The threat is that we could argue in this direction in every case, ending up with justice*, justice**, …, temperance*, temperance**, and so on. We are back where we started. But rather than a proliferation of motivational requirements, we have a proliferation of virtues. This makes it impossible to cope with justification skepticism *de re*.

A second objection is concerned with the first premise rather than the second. If we are being honest, someone might say, when we observe the human form of life, we see that what good hangs on is not the keeping of our promises but the keeping of our promises^{o}. The difference between these two practices is that whereas the former is defined over the domain of all human beings, the latter is defined over some particular social group. It is an Aristotelian necessity that people within the same nation or tribe keep their promises—it is necessary for a human being to

belong to such a community—but not that they keep their promises *simpliciter*. The corresponding inference is:

(1) Human beings keep their promises^{o}.
(2) Being just^{o} is a necessary condition of reliably keeping one's promises^{o}.
(3) Therefore, human beings are just^{o}.

Justice^{o} is similarly restricted. The objections are not mutually exclusive. Consider that our first objector might say to our second objector: what you have proved is not that human beings are just^{o} but that they are just^{o}*. And their spirit is the same: what we can find or infer from experience of what is good for human beings is simply not enough to ground the ethical edifice. My response to these objections will go beyond interpretation of Foot and instead make use of some Thompsonian ideas about studying a form of life.

Let me start with the second objection. I will first reply that even if there is good which hangs on promise-keeping^{o}, there is also good which hangs on promise-keeping. After all, insular societies that don't extend the notion of a contract to outsiders will not be able to trade. Moreover, societies in which the virtues are so restricted will be more closed in general—may be more susceptible to stagnant and entrenched belief systems, the loss that comes with war, genetic drift, and so on. The goodness we attribute to cosmopolitanism seems to find its roots here. If this is the way we understand the objection, I think this is enough to cope with it. But maybe this reply is not one we could make. What our objector might be saying is not that promise-keeping^{o} exhausts the good that would be found in promise-keeping, but that we cannot identify any practice of promise-keeping over and above the various practices of promise-keeping^{o}.[42]

When we form an Aristotelian categorical, we represent the trait in question as being

[42] Michael Thompson says that this is what the world looks like on the Humean theory of justice. He argues that this view is unacceptable because of its intolerable ethical consequences (that we might not have to view other human beings as equals or as having moral status) and this may be just what the objector is accusing me of falling into. See 'What is it to wrong someone?' 374 - 6.

characteristic of the species *as* species. Wherever an army ant is found, it is such that, *in its colony*, its morphology determines its role. So we can say that an army ant's morphology determines its role.º A roleº is different from a role in that it is relativized to an army ant's colony. The account the objector proposes is that wherever a human being is found, they are such that, in their social group, promises are to be kept. But there is a difference here. The colony of an army ant will feature in an account of its natural history *as* a colony. Colonies have a definitive structure, function, and place in the natural history of the organism. Can we say the same thing of the amorphous 'social group'? The problem is that, if we restrict the domain of promise-keeping within the Aristotelian categorical, we need to make sense of how this domain restriction is part of or characteristic of the species, just as the army ant's relationship with its colony will be a relationship characteristic of the species. If we can develop an account of 'social group' that serves this role, then there is a chance for this objection. But my feeling is that it cannot be done: our place in a social group is not biologically programmed in the same way as an army ant's place in its colony is.

The objector might say that a social group takes on a different form in the case of each human being—the city-state for the Athenian, the caliphate for the Medieval muslim, and the "western world" for the modern person. Then we must turn to more abstract characteristics of a social group to see how it is characteristic of a species, such as shared institutions or concepts. Yet this is precisely what is in question: we want to know over which range of relationships the institution of promise-keeping is defined, whether this activity which looks like promise-keeping is the same as this other one found in a different historical context. If recognition of virtues comes first, it may be that our social groups are wider than we first supposed. Perhaps they extend across humanity: promise-keeping is an institution which is good for the species, and if

institutions define social groups, then we are closer to each other than we think.[43]

And this same insight can help us with the first objection, but here I am going to emphasize a different consequence. When understanding some action which has come to be an Aristotelian necessity among human beings, we see it as a characteristic of a naturally good individual human being's life. If it is characteristic of human beings that they have four-chambered hearts, it is a *naturally good* characteristic of this human being that they have a four-chambered heart.[44] We can alter our inference to reflect this. We are not just looking for the reliable production of this action, but what would make this reliable production characteristic of this person:

(1) Human beings keep their promises.
(2) Being just is a necessary condition of being a characteristic promise-keeper.
(3) Therefore, human beings are just.

We're still not there yet. What we are saying when we say (2) is not that this disposition is necessary to have to produce this action reliably, but that this one characteristic underlies this one other characteristic. That is the connection which is being drawn.

So could some other characteristic underlie the characteristic of promise-keeping? Yes, but—and this is my second point—we are drawing this connection within the natural history of the human being. We have to ask: could some other characteristic underlie the characteristic of promise-keeping in human beings? We can exclude the example of being held under social sanctions because this is not the kind of thing which could be a characteristic of the species *qua* species. If some action widespread in a form of life looked as if it were being forced among its members, finding out that it was part of this species's natural history would make one think it was not being

[43] Thompson thinks we can tell a causal story if we are asked what makes this social group a unified social group, but this causal story cannot be characteristic of the species. See 'What is it to wrong someone?' 355-8.
[44] This is similar to Thompson's point that a disposition is nothing more than the presence of a practice in some individual. See *Life and Action.* 208.

forced. If it needed to be forced in a specific case, that would be an exception—it would not be an example of natural goodness.[45] The more difficult case is the one which involves notions like justice* or justice**. Unlike duress, it is clear that these are the kinds of things which could be characteristics of the species. But although justice* could be a characteristic, it is *not.* What I mean is that there is no characteristic of human beings which is an instance of justice*. Just as having some set of undetached rabbit parts is necessary for some thing to grow and develop as a rabbit, so it may be that justice* is necessary to be a characteristic promise-keeper. But the way a rabbit has those undetached rabbit parts is by *being a rabbit*. The way a human being is just* is by being just.

The point of our inference is not discovery of justice as something on which good hangs for human beings, but the discovery that good hangs on justice, which we know to exist or can identify. The inference is of the same kind as the inference the medical scientist makes when, upon learning that the gut microbiome retreats to the appendix upon infection, recognizes that the appendix is not a vestigial structure. A scientist might infer the existence of some biological structure or process based on phenomena, just as someone might infer the existence of justice* from the Aristotelian necessity of promise-keeping, but that will not be the discovery of that phenomena. In fact, in the case that interests us, there is no characteristic which we could call justice* if we were to look for it.

And just as the medical scientist's investigation teaches us that a healthy person has an appendix by showing that good hangs on it, Foot's argument will help us see that a good person is just by showing that good hangs on it. On that note, here is the final form our inference will take:

[45] Consider: the bird that drops its young out of the nest so that they learn to fly, intraspecific competition for mates, hierarchies among elephant seals.

(1) Keeping promises is characteristic of the human form of life.
(2) In the human form of life, being just is a necessary condition of being a characteristic promise-keeper.
(3) Therefore, human beings are just.

I think this is right.

Solutions

We are now in a position to solve the problems that were raised in the beginning of this essay. What we want to know first is how this account of virtue will allow Foot to meet the Apprehension Requirement without opening herself up to justification skepticism read de re. Rather than say that people must recognize particular motivational requirements, we can say that the connection between a person's good intentions and their good action is mediated by virtues.

The key here is that the Aristotelian necessity of a virtue can be derived from some small set of empirical observations and that this virtue has wider application than the situations from which it was derived.[46] The first point in favor of justification skepticism *de re* was that empirical investigation would not reveal the goodness and badness of some acts. But now this does not matter, for empirical investigation is how one comes to know the virtues, and possession of the virtues is how one comes to know the goodness and badness of actions. The second point in favor of this skepticism was that this could not be how people learn right and wrong. The account I have given, on the other hand, seems to capture this perfectly. Ostention happens in some cases, and from there, we learn patterns in the moral sphere.

Here is an invented story about how that might happen. I recognize the truth of the Aristotelian categorical that human beings care for their children. From here I am able to recognize the Aristotelian necessity of benevolence, a characteristic which disposes me to act in certain ways in

[46] In the example I have used so far, the Aristotelian necessity of justice was inferred from promise-keeping. But we do not need to limit ourselves this much, that is not what meeting the Apprehension Requirement asks of us. It only requires plausibility.

a wide variety of situations; this is the indirect practical recognition of motivational requirements I would otherwise have to acknowledge. I cultivate the virtue of benevolence on these grounds. Now suppose I find someone who is not my kin in distress. In the light of the virtue of benevolence, a motivational requirement becomes salient: if someone is in distress, I ought to help them on that basis.[47] My understanding of what is good to do in this example is not in the particular recognition of this motivational requirement, but in the recognition of the virtue that disposes me to recognize the motivational requirement in action. So in the case of the person who thinks "I can't take that, it's hers", they may also think "I can't take that, it's unjust."

Before moving onto how this helps us with objectivity skepticism, my main aim, I should explain how we have defused practicality skepticism. Foot's account connects moral judgments about particular actions to practical rationality through the concept of a virtue in the way I have been articulating in order to show that such judgments "serve to produce and prevent action."[48] In outlining the way that Foot can meet the Apprehension Requirement, we also outline precisely where a person's reasons can come from in acting virtuously. And since rational considerations can move us to action, practicality skepticism cannot be successful.

[47] The kind of salience involved here does not need to take the form of conscious thought. It may simply be a way of perceiving the situation. See Butler, 'Character Traits in Explanation,' 220 - 221.
[48] Foot, *Natural Goodness*, 9.

Works Cited

Butler, Douglas. "Character Traits in Explanation." *Philosophy and Phenomenological Research* 49, no. 2 (1988): 215-238.

Foot, Philippa. *Natural Goodness*. Oxford: Oxford University Press, 2001.

Foot, Philippa. "Rationality and Goodness." *Royal Institute of Philosophy* 54 (2004): 1-13.

Haase, Matthias. 'Practically Self-Conscious Life.' *Philippa Foot on Goodness and Virtue.* Edited by John Hacker-Wright. London: Palgrave-Macmillan. 2018.

Korsgaard, Christine. 'Constitutivism and the Virtues.' *Philosophical Explorations* 22 (2019): 98-116.

McDowell, John. "Two Sorts of Naturalism." *Virtues and Reasons: Philippa Foot and Moral Theory.* Edited by Rosalind Hursthouse. Oxford: Clarendon Press. 1995.

Muller, Anselm. '"Why Should I?" Can Foot Convince the Skeptic?' *Philippa Foot on Goodness and Virtue.* Edited by John Hacker-Wright. London: Palgrave-Macmillan. 2018.

Thompson, Michael. 'Apprehending Human Form.' *Royal Institute of Philosophy* 54 (2004): 47-74.

Thompson, Michael. *Life and Action*. Cambridge: Harvard University Press. 2008.

Thompson, Michael. "What is it to Wrong Someone? A Puzzle about Justice." *Reason and Value: Themes from the Moral Philosophy of Joseph Raz. Edited by R. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith.* Oxford: Oxford University Press. 2004.

Ostensibly Competing Kantian Duties and the Limits of Violent Self-Defense

Sofia Stutz — *Northwestern University*

Introduction

In Kantian moral theory, the Categorical Imperative (CI) is the supreme principle of morality that unconditionally binds all agents to moral conduct irrespective of human desires or anticipated outcomes of action. One formulation of the CI is the Formula of Universal Law (FUL), which commands us to "Act only according to that maxim through which you can at the same time will that it become a universal law" (G 4:421). Roughly, this formula asks us to imagine a world in which the act in question—formulated as a maxim, or a moral principle—functioned as a universal law. If this universalization results in a contradiction, the act is impermissible. Kantian scholars contest the sense in which a maxim contradicts itself upon universalization but I will not explore that here. What I'd like to examine is the surprising fact that standard interpretations of the FUL remain ill-equipped to reject all violent actions as strictly prohibited.[1] [1]In the following paper, I survey Christine Korsgaard's Practical Contradiction Interpretation (PCI) in particular as it is the leading standard approach.[2] Next, I introduce Donald

[1] Donald Wilson, "Murder and Violence in Kantian Ethics," in *Natur und Freiheit. Akten des Xii. Internationalen Kant-Kongresses*, ed. Violetta L. Waibel, Margit Ruffing and David Wagner (De Gruyter, 2019), 1.
[2] Barbara Herman's approach in "Murder and Mayhem: Violence and Kantian Casuistry" is also standard in its treatment of violent actions; she categorizes some but not all violent actions as perfectly prohibited. Using what I call an agency-focused lens, she locates the immorality of murder in its devaluation of rational agency (425). On her view maxims that generate a Contradiction in Conception (CC) fail universalization because they are incompatible with the integrity of the will. Upholding this integrity of the will—both in ourselves and in others—means respecting its capacity for self-direction, for autonomy. Consequently, maxims that prohibit people from acting for their own reasons are impermissible. Maxims of coercion and deception, for example, generate a CC because they form a direct attack on the will (424). Both of these maxims fail the universalization condition because they allow for the manipulation of one agent's desires and beliefs for the purpose of another; they convert one person's agency

Wilson's interpretation of the FUL, which departs from standard accounts in its satisfactory classification of violence as strictly prohibited. Drawing attention to the absolutist pacifist principle which seems to emerge from his account, however, I examine the questionable rigorism of the prohibition. It appears to reject the use of violence even in self-defense. And this seems to be at odds with other Kantian principles, namely, the duty of self-preservation, and the authorization to use coercion when others interfere with our freedom. We are ostensibly torn by two competing duties: the duty prohibiting violence, and the duty to defend ourselves. I resolve this puzzle by dispelling its status as a problem to be solved in the first place. The authorization to use coercion amounts to the permissibility of defensive violence, and this is consistent with moral rules, not opposed to them. The real question, I argue, is not how to override the duty prohibiting violence in favor of self-defense, but how to identify the appropriate amount of force to use in a defensive response. While I don't offer a complete account of the solution here, I gesture at the development of one.

---

into the means of another's end. And it is contradictory to rationally conceive a self-destructive universal law (423). Maxims that generate a Contradiction in the Will (CW), on the other hand, occur when we will universal laws that conflict with our ability to exercise our agency effectively. These maxims are wrong because they interfere with the conditions of rational agency. Non-coercive violent acts, for example, attack the "material condition of human agency"— the body. And for humans, the body is a necessary condition of rational agency. Maxims of non-coercive violence generate a CW rather than a CC because they attack the material condition of agency rather than the agency itself (424). If I kill you, for example, I am not attacking your will per se, but your entire existence. I am not directly controlling your will but, in taking your life, eliminating the possibility of you exercising your will at all.
Sensitive to the complaint that a prohibition against violence should be a perfect rather than imperfect duty, Herman expresses surprise about the result of her CI procedure but nevertheless defends her account by insisting that we can both classify duties prohibiting non-coercive violence along with duties of beneficence and simultaneously maintain a distinction between them (421). These duties belong together, she argues, because they both concern themselves with the necessary conditions of rational agency. These imperfect duties differ, however, because they discount different conditions of agency; non-coercive violence discounts the condition of vulnerability while nonbeneficence discounts the condition of need. As rational agents, we can neither exempt ourselves from our dependence on others for help nor can we escape our bodily vulnerability to death since we are mortal creatures. Their overarching similarity yet descriptive difference suggests that these obligations can coexist as imperfect duties, or so she argues. But her account isn't entirely convincing. In addition to unsatisfactorily classifying non-coercive violence as a violation of an imperfect duty rather than a perfect one, and overlooking the important differences between coercive violence and other perfect prohibitions, Herman's account fails to sufficiently address concerns about the latitude of imperfect duties. See Donald Wilson's article "Murder and Violence in Kantian Ethics" for a full critique.

The Formula of Humanity Implies a Stringent Prohibition Against Violence

Like the FUL, The Formula of Humanity (FH) is a formulation of Kant's CI, and it runs: "So act that you use humanity, in your own person as well as in the person of any other, always at the same time as an end, never merely as a means" (G 4:429). Human value, according to Kant, is different from the value of objects. Humans have a *dignity*, or irreplaceability, stemming from their capacity to choose and pursue ends. This irreplaceable value of human life differs from that of objects, which have a *price* in virtue of their replaceability. Though objects may be irreplaceable in the sense that, say, a particular vase is impossible to replace, they nevertheless bear an exchange value. Nothing can compensate, however, for the loss of a person. Given the FH, which categorically commands us to treat others with respect, it seems straightforward to conclude that there exists an absolute prohibition against violence, which necessarily fails to treat human beings as ends in themselves. But in what sense does violence fail to do this? I believe it does so in the following ways: to kill someone is to disregard the irreplaceability of human life, to violently injure someone for the sake of expediency is to treat that person as a mere means to an end, and to inflict violence for its own sake is to take pleasure in the suffering of others, which, in turn, means treating others as an instrument to one's own pleasure. But, as I mentioned above, standard interpretations of the FUL don't yield a perfectly stringent prohibition against violence.

But what exactly does a perfectly stringent prohibition against an act entail? For Kant, it means that a maxim—the action you are evaluating formulated as a moral principle—creates a *Contradiction in Conception* (CC) upon universalization. This procedure helps us identify duties of the most stringent variety, called *perfect duties*, which require or, more often, forbid particular actions. Maxims that produce a CC "cannot even be *thought* without contradiction as a universal law of nature" (G 4:424). In other words, universalizing the maxim in question would produce some sort of internal contradiction that makes the action out as impermissible. Different

interpretations understand the nature of the contradiction differently. Here I will briefly survey Korsgaard's Practical Contradiction Interpretation and examine its limitations with regard to violent maxims.

Under the PCI, an action is impermissible if its maxim becomes self-defeating upon universalization. In other words, the efficacy of the method you employ for your purpose would fail in a world where everyone used that method to accomplish the very same purpose. Willing such a maxim results in a contradiction because, in effect, you "will the thwarting of your own purpose."[3] Take Kant's example of the false promise: I ask you for some cash, promising to pay it back, with no intention of doing so. Converting this into a maxim, we get: people who need cash make false promises in order to get that cash. To carry out the PCI, we must imagine that the act in question, taken for a particular purpose, becomes the standard universal method for achieving that purpose.[4] If the method continues to be efficacious upon universalization, the maxim is permissible. If the method simply could not *work* if universally practiced, the maxim is impermissible. So in our example, we must imagine that making a false promise in order to get ready cash is the standard method for doing so, and then evaluate its effectiveness in the world of the universalized maxim. As it turns out, the false promise method, upon universalization, ceases to be effective; you could not successfully make a false promise to get cash in a world where everyone did that because no one would believe you. In the words of Kant, "no one would believe he was being promised anything, but would laugh about any such utterance, as a vain pretense" (G 4:422).

Established practices depend for their existence on people's adherence to the rules for the majority of the time. For this reason, people only rely on the practice of promising—in other

[3] Christine Korsgaard, "Kant's Formula of Universal Law," in Creating the Kingdom of Ends, (Cambridge; New York, NY, USA: Cambridge University Press, 1996), 78.
[4] Ibid., 92.

words, this practice exists—because most agents comply with the rules of the practice and abide by their promises. Otherwise, the practice would die off. The PCI exposes unfairness by singling out actions that depend for their efficacy on the agent making an exception for herself. In the case of the false promise, "other people's honesty makes your deceit effective."[5]

The PCI works particularly well with conventional wrongs, as opposed to natural wrongs.[6] Acts that rely on natural laws of nature, like murder, belong to the latter category, while acts that depend upon general adherence to a practice, like promising, belong to the former. While it easily classifies conventional wrongs as strictly forbidden, the interpretation only blocks *some* natural wrongs in the CC. Namely, the kinds of natural actions whose efficacy depend upon their exceptional use.[7] To illustrate this point, Korsgaard provides the example of the murderous employee. Imagine "you are second in line for a job, and are considering murder as a way of dealing with your more successful rival."[8] The maxim would be: in order to secure a job for which one is runner-up, one kills the employee first in line. This maxim fails the universalization condition because if every worker took up this method for securing a job, then by domino effect, each worker would be killed by the one next in line, and the agent proposing the maxim would himself become the victim of his method upon universalization. Being alive is a necessary condition for securing a job, and murder impedes its victim from fulfilling this condition. So the murderous employee's method backfires on him, failing to secure him a job.

Clearly, the PCI does manage to rule out some violent maxims. However, it fails to rule out violent maxims that don't depend for their efficacy on their exceptional use. Indeed, Korsgaard observes that if the purpose of a violent act is simply "getting someone dead," the PCI does not appear to reject it since the method of killing in order to get someone dead doesn't lose

[5] Korsgaard, "Kant's Formula," 93.
[6] Ibid., 97.
[7] Ibid.
[8] Ibid., 98.

its efficacy upon universalization.[9] Nevertheless, Korsgaard insists that the PCI can ward off the universalization of violent methods if we frame maxims the following way:

> [I]f we include as part of the purpose that the agent wants to be secure in the possession of an end, we can get a practical contradiction in the universalization of violent methods.[10]

The universal use of violent natural means would undermine everyone's security, upon which all our possessions, including our ends, rest.[11] Rational agents need security in order to protect their possessions and pursue their ends. For this reason, violent methods cannot be universalized.

While protecting both one's ends and possessions requires security, it seems strange to classify an end as a possession. Moreover, the PCI as a whole hinges upon the unfairness of making exceptions for yourself—and this focus fails to capture what is distinctively wrong about murder and other violent acts. As Wilson explains, the PCI "classifies at least some" acts of violence and murder as contrary to perfect duties but "it does so on the basis that agents proposing to use murder to further their ends could not hope to do so securely in a world in which everyone did the same."[12] But murder isn't wrong because it's unfair for me to commit it when everyone else refrains from doing so.[13] This reason seems "off." We should reject maxims for reasons that explain what is objectionable about them.[14]

On standard accounts like the PCI, maxims of violence that aren't blocked through CC get blocked instead through the *Contradiction in the Will* test. Under this procedure, a maxim is impermissible if a contradiction results from *willing* it as a universal law (G 4:424). To will an

[9] Korsgaard, "Kant's Formula," 98.
[10] Ibid., 99
[11] Ibid.
[12] Wilson, "Murder and Violence," 1.
[13] Ibid.
[14] Barbara Herman, "Murder and Mayhem: Violence and Kantian Casuistry." *The Monist* 72, no. 3 (1989): 413, https://doi.org/10.5840/monist198972321.

end is to actively pursue and commit yourself to it. If you find that you *can* rationally will a maxim as a universal law, then you have merely identified a permissible maxim. One example might be: "I will have cereal for breakfast, in order to start my day with something nutritious and easy to prepare."[15] If you find that you *cannot* rationally will a maxim as a universal law, on the other hand, then the maxim produces a CW, and you should adopt the inverse of the maxim in question. If, for instance, we universalize the maxim of nonbeneficence—of never helping others—we find that we cannot rationally will it. Given our inherent dependence on other human beings, we cannot rationally will a world in which nobody ever offers a helping hand, since no agent, in such a world, could successfully pursue her ends. And since we cannot rationally will the maxim of nonbeneficence, we must adopt its inverse: the duty of beneficence, an imperfect duty. Unlike perfect duties, which require or forbid particular actions, imperfect duties allow for latitude in choosing how and when to fulfill them; they are ongoing commitments, such that only some pattern of action or some attitude toward an end would count as a violation of the duty—not some particular action or omission. In the case of beneficence, how and when to help others is to some degree, up to us, but never helping others is objectionable.

But the prohibition against violence should not be an imperfect duty; it should leave no room for discretion in choosing how and when to fulfill it—it should simply forbid it. So my first objective is to find an interpretation of the FUL that blocks violent maxims through CC. It would seem that only an interpretation that does this can properly uphold the basic commitment outlined in the Formula of Humanity to always treat others as ends in themselves.

[15] Thank you to Kyla Ebels-Duggan for suggesting this example.

Wilson's Alternative Account

Unlike standard approaches, Wilson's account yields a stringent prohibition against violence. Wilson uses Kant's discussion of suicide as a model for understanding violence directed at others. In the *Metaphysics of Morals*, Kant classifies the prohibition against suicide as a perfect duty. He does so on the grounds that suicide deprives one of "one's *capacity* for the natural (and so indirectly for the moral) use of one's powers" (MM 6:421). Namely, the power that lies in and can only be realized through the exercise of our rational natures—our capacity for self-direction.[16] According to Kant, perfect duties require respect for the basic conditions of rational agency. Actions that violate these duties rob "us of the capacity to use our powers in the service of rational self-constraint."[17] Imperfect duties, on the other hand, require "respect for the conditions necessary for the effective exercise of rational agency."[18] The former duty deals with moral health and the latter, with moral prosperity (MM 6:419). The prohibition against suicide revolves around a preoccupation with moral health—with respecting our rational capacity, which depends upon our physical existence.

Wilson argues that Kant's account of suicide applies to cases of violence against others. Killing another, he claims, is wrong for similar reasons to suicide. The immorality of suicide lies in the "failure to respect the integrity and proper functioning of our bodies."[19] And the body, claims Wilson, is a necessary condition of rational agency. Thus, violence against both the self and the other is a perfect prohibition.

[16] Wilson, "Murder and Violence," 5.
[17] Ibid.
[18] Ibid.
[19] Ibid.

An Absolutist Pacifist Principle

Wilson's account implies the existence of an absolutist pacifist principle; given the perfect obligation to refrain from violence, it would seem to follow that violence is *never* permissible. But is self-defense impermissible? The pacifist principle does not seem to distinguish defensive from assaultive violence. It appears to be a blanket prohibition against all violence. Suppose I start punching you in the ribs and you start to feel some bones cracking. Are you barred from defending yourself violently because of the strict prohibition on violence? How are you to respond in the face of my assaultive violence? It seems unduly harsh to claim that you have absolutely no right to defend yourself through violent means.

Wilson claims that violence amounts to a failure to respect the body—a necessary condition of rational agency. By punching you, I disrespect the integrity of your body and thus, your rational agency. According to the absolutist pacifist principle, it would seem that if you started punching me in response, you would be guilty of the very same charge. By punching back, you disrespect the integrity of *my* body and rational agency. But this makes it seem like a perfectly symmetrical dynamic. I was the one who started *attacking* you and you were attempting to defend yourself. The pacifist principle as such categorizes all forms of violence as equally impermissible, failing to take into account the assaultive/ defensive dynamic of certain cases, which we might intuitively think has some degree of relevance in assessing their moral quality.

The questions above illustrate a skepticism toward the uncompromising nature of the absolutist pacifist principle. In core cases—ordinary, everyday moral dilemmas—the principle seems attractive; resolving our everyday problems through non-violent means is reasonable and laudable. In radical cases of violent attack, however, the absolutist principle can seem

unreasonable.[20] For it forbids the use violence in self-defense, and this seems to be at odds with another principle in Kantian moral theory that indicates we do, in fact, have a duty to protect ourselves.

The Authorization to Use Coercion

In *The Metaphysics of Morals,* Kant lays out the duty of self-preservation, claiming that "The first, though not the principal, duty of man to himself as an animal being is *to preserve himself* in his animal nature" (MM 6:421). But the "need to maintain" the body "in itself cannot establish a duty" (MM 6:446). Actions that violate the duty of self-preservation are wrong because they undermine or destroy our rational agency.[21] It is for the protection of our rational capacity—which grounds human dignity for Kant—that we are obligated to preserve our bodies. But what can this duty tell us about self-defense? Kant explicitly claims that he means only to highlight what the duty of self-preservation prohibits us from doing, not what it gives us license to do.[22] The chapter on duties to the self "deals only with negative duties and so with duties of omission," he writes (MM 6:421).

In the Private Right section of *The Metaphysics of Morals*, however, Kant outlines a stronger permission to defend ourselves—the authorization to use coercion. This positive view, which I will explain shortly, relies on his conception of freedom. For Kant, the right to freedom is our only innate right, and it accords us "independence from being constrained by another's choice, insofar as it can coexist with the freedom of every other in accordance with a universal law" (MM

[20] And this concern extends well beyond the prohibition against violence to other strict prohibitions such as deception, as the controversy surrounding Kant's essay "On the Supposed Right to Lie" reveals. See Tamar Schapiro's "Kantian Rigorism and Mitigating Circumstances" for an approach to the rigorism question as it relates to lying, and Michael Cholbi's "The Constitutive Approach to Kantian Rigorism" for a critique of Schapiro's approach.
[21] Michael Cholbi, "The Murderer at the Door: What Kant Should Have Said," *Philosophy and Phenomenological Research* 79, no. 1 (2009): 24, https://doi.org/10.1111/j.1933-1592.2009.00265.x.
[22] Ibid., 25.

6:238). In other words, the only limit to my use of freedom is your same right to freedom; I can only employ my right to freedom in a way that allows you to exercise your right as well. Conversely, you are to constrain your own use of freedom in view of mine. We limit each other's freedom reciprocally.

With this innate right in mind, Kant formulates the *Universal Principle of Right*, according to which "Any action is *right* if it can coexist with everyone's freedom in accordance with a universal law, or if on its maxim the freedom of choice of each can coexist with everyone's freedom in accordance with a universal law" (MM 6:231). Put differently, actions are right if they are compatible with everyone else's rightful use of freedom. The sort of right Kant refers to here is not moral but political.[23] This kind of right, he argues, is connected with an *authorization to use coercion* (MM 6:231). As long as our actions are compatible with other people's freedom, Kant posits, we may coerce others to respect our use of freedom.[24]

Kant does not, in this context, "understand coercion primarily in terms of the making and carrying out of threats, but instead in terms of reciprocal limits on freedom."[25] And it is not, as some have misinterpreted, a right to punish those who violate one's freedom,[26] but a right to compel perpetrators to refrain from encroaching upon it. So if I hinder you from pursuing your ends, which you do so without violating anyone's freedom, I *wrong* you, and this gives you license to coerce me into respecting your freedom. Your action would then be, as Kant describes it, a "hindering of a hindrance to freedom" (MM 6:231). Arthur Ripstein explains that a hindrance to a hindrance of freedom

> is not a second wrong that mysteriously makes a right, because the use of force is

[23] Kyla Ebels-Duggan, "Kant's Political Philosophy," *Philosophy Compass* 7, no. 12 (2012): 897, https://doi.org/10.1111/j.1747-9991.2012.00525.x.
[24] Ibid.
[25] Arthur Ripstein, *Force and Freedom: Kant's Legal and Political Philosophy* (Cambridge, Mass: Harvard University Press, 2009), 301.
[26] Ibid., 54.

> only wrongful if inconsistent with reciprocal limits on freedom. So force that restores freedom is just the restoration of the original right.

Japa Pallikkathayil also puts it nicely:

> actions needed to thwart a rights violation only prevent the would-be violator from performing an action that is not in her discretionary sphere. Hence such defensive actions are consistent with the violator's equal external freedom.[27]

In short, your hindrance to my hindrance of your freedom is consistent with the only limitation on freedom—mutual respect for each other's use of it. I believe that we can interpret this authorization to use coercion as a permission to defend ourselves against violence. Stronger than the negative duty of self-preservation, this positive principle permits the use of defensive violence.

For Kant, the authorization to use coercion appears to be consistent with his moral philosophy. This is clear from the relationship he outlines between equal external freedom and the value of humanity; respect for the former is ultimately grounded in respect for the latter.[28] According to Kant, all people have an innate right to freedom in virtue of their humanity.[29] And we have humanity in virtue of our rational capacity—our ability to autonomously choose and pursue ends. This rational capacity is what makes humans ends in themselves rather than mere means. So when a perpetrator violates another's freedom to pursue these ends, he limits her capacity to engage in autonomous action, and in this sense, treats her as a mere means.[30] By disregarding her freedom, he interacts with her as if she were a mere object.[31] Indeed, according to Kant, "it is clear that the transgressor of the rights of human beings is disposed to make use of the person of others merely as a means" (G 4:430). Seen as a justified response to the human right

[27] Japa Pallikkathayil, "Deriving Morality from Politics: Rethinking the Formula of Humanity," *Ethics* 121, no. 1 (2010): 134, https://doi.org/10.1086/656041.
[28] Ibid., 130.
[29] Ibid., 133.
[30] Ibid.
[31] Ibid., 142.

to freedom, and thus, a recognition of humanity, the authorization to use coercion upholds the value of humanity, which lies at the heart of Kant's moral theory.

Given the compatibility between the authorization to use coercion and the value of humanity, there doesn't seem to be a need to make an exception in Kant's moral theory for self-defense in the first place; the permission to defend ourselves need not *override* the perfect prohibition against violence because these principles don't compete at all. The (schematic) task I turn to next is that of identifying what violent self-defense consists of and what limitations our duties to others place on it.

Identifying a Threshold

We've addressed the question of *whether* violence in self-defense is appropriate but this does not give us any insight into the question of *how much* violence in self-defense is appropriate. Those who hinder others' rightful use of freedom through violence commit wrongdoing but are not, for this reason, "beyond the pale" of the moral community; no action, I believe, can eliminate one's humanity.[32] The "anything goes" approach to self-defense runs the risk of denying the perpetrator's humanity. So we need some sort of standard of limitation on defensive violence.

For Kant, the authorization to use coercion extends only to defensive action necessary for thwarting the violation of one's rights.[33] "If pushing you away would keep you from my apple," for instance, "I am not entitled to cut off your hand." [34] Defensive action that goes beyond necessary force violates the perpetrator's innate right to freedom. On the surface, it seems as though we have located an appropriate standard of defensive violence, something we might call the *principle of necessary force*: If A violates B's rightful use of freedom, B will apply

[32] Trudy Govier, "Forgiveness and the Unforgivable," *American Philosophical Quarterly* 36, no. 1 (1999): 62.
[33] Pallikkathayil, "Deriving Morality," 134.

only the amount of force necessary to restore B's freedom. But it is not clear how one could apply this principle with complete certainty in one's methods. How can one possibly identify the exact amount of force necessary to restore one's freedom, and which specific actions are best suited to doing so?

I believe this task—that of adjudicating the limits of defensive action—lies in the realm of political philosophy. According to Kant, it is only possible to respect every person's innate right to freedom through the joint agency or general will of a people since no individual has the legitimate authority to enforce or establish the boundaries of private rights.[36] To act jointly as a public just is to act as a state.[37] So the only legitimate source of authority for adjudicating the limits of self-defense is the state. The legislative function of the state, which specifies the law, the executive function, which enforces the law, and the judicial function, which settles disputes between individuals over the law, collectively comprise the omnilateral will of the people.[38]In the case of self-defense, the legislative branch takes up the task of defining appropriate limits of defensive violence, while the other two branches enforce and interpret its limits. The details surrounding the proper function of these branches are beyond the scope of this essay.

Conclusion

I began with the surprising fact that standard interpretations of the Formula of Universal Law fail to classify all violent actions as strictly prohibited. Given the Formula of Humanity's command to respect the dignity of persons, I reasoned, any convincing interpretation of the Formula of Universal Law should create a perfect duty forbidding violence. Identifying Wilson's account as a plausible contender, I then drew attention to the absolutist pacifist principle which emerged from his account. This principle, I claimed, was too rigorous, for it denied the use of violence in self-defense. And according to Kant, we have a duty of self-

preservation. As a negative rather than positive duty, however, this duty of self-preservation isn't a strong enough permission to defend ourselves. Shifting my focus to the political section of the *Metaphysics of Morals*, I examined Kant's authorization to use coercion, and in view of its positive nature, interpreted it as permission to use defensive violence. Here we appeared to run into a problem: the duty prohibiting violence and the permission to defend ourselves seemed to compete. Outlining the consistency of the authorization to use coercion with Kant's moral framework, however, I concluded that the ostensibly competing duties did not compete at all. Finally, I turned to a discussion of the appropriate threshold for defensive violence, and claimed that within Kant's paradigm, only the general will—manifested in the state—is legitimately positioned to adjudicate the precise limits of self-defense.

Works Cited

Cholbi, Michael. "The Murderer at the Door: What Kant Should Have Said." Philosophy and Phenomenological Research 79, no. 1 (2009): 17–46. https://doi.org/10.1111/j.1933-1592.2009.00265.x.

Ebels-Duggan, Kyla. "Kant's Political Philosophy." Philosophy Compass 7, no. 12 (2012): 896–909. https://doi.org/10.1111/j.1747-9991.2012.00525.x.

Govier, Trudy. "Forgiveness and the Unforgivable (Moral Philosophy)." American Philosophical Quarterly 36, no. 1 (1999): 59–75.

Herman, Barbara. "Murder and Mayhem: Violence and Kantian Casuistry." The Monist 72, no. 3 (1989): 411–31. https://doi.org/10.5840/monist198972321.

Kant, Immanuel. Groundwork of the Metaphysics of Morals. Translated by Mary J. Gregor. Cambridge, U.K.; New York: Cambridge University Press, 1998.

Kant, Immanuel. The Metaphysics of Morals. Translated by Mary J. Gregor. Cambridge: Cambridge University Press, 1991.

Korsgaard, Christine. "Kant's Formula of Universal Law." In Creating the Kingdom of Ends, 77-105. Cambridge; New York, NY, USA: Cambridge University Press, 1996. Pallikkathayil, Japa. "Deriving Morality from Politics: Rethinking the Formula of Humanity." Ethics 121, no. 1 (2010): 116–47. https://doi.org/10.1086/656041.

Ripstein, Arthur. Force and Freedom: Kant's Legal and Political Philosophy. Cambridge, Mass: Harvard University Press, 2009.

Wilson, Donald. "Murder and Violence in Kantian Ethics." In Natur und Freiheit. Akten des Xii. Internationalen Kant-Kongresses., edited by Violetta L. Waibel, Margit Ruffing and David Wagner, 2257-64. De Gruyter, 2019.

An Antirealist Account of Gender as a Conceptual Relationship

Sophia Heimbrock — *New York University*

Introduction

In the past few decades, gender — more specifically, what gender really is — has become prominent in mainstream cultural discussion, in part due to the increased visibility of transgender people. Gender non-conforming people have always existed, many of them publicly, but it was not until recently that phrases like "I *am* a woman" or "I *am* non-binary" were commonly used by non-cisgender and gender non-conforming people.[1] Phrases like these are intended to convey that not only does one feel like a certain gender, but one *is* a full-fledged member of that gender. This emerging landscape of gender identity calls for careful consideration of our gender vocabulary — more specifically, what it is for someone to say that they *are* a certain gender.

My account of gender, to borrow from Sally Haslanger, is a critical descriptive one: I aim to determine what our gender vocabulary tracks, i.e., what it is to call oneself a given gender and what it is to be that gender. I will argue that someone is the gender that they are in virtue of their relationship to the gender concept that is assigned to them. Moreover, I will argue that personal gender statements (e.g., the statement "I am a man") ought to be understood as an expression of this conceptual relationship, rather than a self-conferral of a real identity.

[1] Such phrases are not only common but are now standard. To say that a trans woman "feels like" or is "living as" a woman is often met with resistance; the proper terminology is considered by most trans people (and progressives) to be that she simply "is" a woman.

In defending this view, it will be useful to consider two prominent alternative views on gender: one, proposed by Sally Haslanger, that gender is a social class; and another, by Michael Rea, that gender is a real self-conferred identity. First, in §1, I will consider Haslanger's view and reject it on the grounds that it does not apply to transgender individuals and is therefore not an adequate account of gender.[2] In §2, I will agree with Rea that gender is "an interpretation of one's own inner experience"; however, I will argue that Rea is mistaken in thinking of gender as a real identity. In §3, I will expand on my own view and argue that someone is the gender that they are in virtue of their internal relationship to the gender concept that is externally conferred on them based on their assigned sex. In §4, I address a potential objection that my view is merely a variation on dispositionalism. In §5, I'll examine some consequences of my account, and in §6 I will conclude by emphasizing its practical advantages.

§1. Haslanger and transgender individuals

Haslanger argues for an account of gender as an externally conferred social class (Haslanger 2000). On her view, someone is a woman if and only if a) they are perceived to have female reproductive anatomy; b) their being perceived to have this anatomy marks them within their social context as someone who ought to be oppressed; and c) they are actually oppressed due to fulfilling (a) and (b). The inverse constitutes what it is to be a man. This view prima facie excludes non-passing transgender individuals — a trans woman who is not perceived to have female anatomy is not a woman, according to Haslanger. What is she, then? In an attempt to account for intersectionality of identity, e.g., black men who are emasculated by violence,

[2] In this paper, I will use the term "transgender" to refer broadly to all people who make personal gender statements that express rejection of the gender concept assigned to them at birth. This includes people who call themselves non-binary. When I refer to a transgender person who expresses rejection by calling themselves the gender that is "opposite" their assigned gender concept (i.e., a "binary" trans person), I will use the term "trans man" or "trans woman."

Haslanger presents a tweaked definition: someone may *function* as a woman in a context Y if she fulfills conditions (a), (b), and (c) above in context Y. Moreover, if she does *not* fulfill those conditions in a context Z, she does not function as a woman in context Z.

Does this notion of a functioning gender work for a non-passing trans woman? We might be tempted to say that a non-passing trans woman functions as a man in many contexts. But we would expect that, though she might be observed to have male anatomy and therefore marked for a privileged position, fulfilling conditions (a) and (b), she will actually be oppressed, not privileged, because she violates masculine gender norms. Therefore, she won't fulfill condition (c) and cannot be described as functioning as a man. On Haslanger's view, then, many transgender men and women can't be described as being men and women, nor functioning as men and women. They are left genderless; on Haslanger's view, for these people to state "I am a woman" conveys no meaning. Since I am operating under the (hopefully intuitive) assumption that personal gender statements like "I am a woman" are always meaningful, this is an unacceptable consequence. Moreover, even if you are not convinced that personal gender statements are meaningful, Haslanger's account of gender clearly fails to describe a demographic of significant size, i.e., transgender people. It should be intuitive, then, that her view is inadequate as a descriptive account of gender.[3]

However, I think a correct implication of Haslanger's view is that in virtue of being perceived to have certain anatomy, one is identified as having a certain set of gender norms apply

[3] Haslanger claims to be providing a "critical analytical" account aimed toward elimination of sexual injustice, so she may argue that her view needn't describe everyone. However, she spends a good deal of time in her paper tweaking her definition of gender to include intersectional identities, so it seems she's aiming for at least some degree of descriptiveness. Moreover, we must ask, isn't sexual injustice perpetrated against transgender people as well? Surely, then, any critical account must include trans people.

to them.[4] I want to take this one step further, borrowing from Ásta's view,[5] and say that simply in virtue of having certain anatomy, an individual has a gender *concept* applied to them (Ásta 2011). I will expand on this at the beginning of §3.

§2. Rea and self-conferred identity

Rea draws a distinction between social gender — the collective representation of a person's identity held by a group of other people — and autobiographical gender — a person's self-authored representation of her identity (Rea 2022). A person's autobiographical gender, which she confers on herself, is her "real" gender because she has privileged epistemic access to her internal experiences. The act of self-conferral of gender is the act of representing oneself as having one's autobiographical gender.[6] To motivate this idea, Rea argues that there are some identities where one's own choices are most salient to the determination of the identity — for example, the identities of being an atheist and being a Star Wars fan. These are the kinds of identities that are self-conferred, on Rea's view.

Being an atheist and being a Star Wars fan, however, are identities constituted of clearly defined attributes (e.g., disbelief in any god and enjoyment of Star Wars). There are times, of course, when the speaker of the statement "I am a Star Wars fan" is met with disbelief (e.g., "You're not *really* a Star Wars fan if you prefer the prequels"). But this disbelief is not based on disagreement over the core definition of being a Star Wars fan — it's based on disagreement over

[4] Given that there are only two sex categories that people are generally sorted into, there are only two sets of gender norms: male and female. The category "intersex" is slowly becoming accepted by some medical researchers and scientists as a third sex category, but given the continued prevalence of forced "corrective" sex surgery on intersex infants and children, I do not think we can yet call "intersex" a fully fledged sex category.

[5] Published under Ásta Sveinsdóttir, but referred to in this paper as Ásta per current publishing convention.

[6] This representation, on Rea's view, need not be public. A closeted trans woman, for example, may represent herself as a woman to herself only; this is sufficient to fulfill the definition of self-conferral.

fringe elements of the definition. Gender identities, on the other hand, are not constituted of clearly defined attributes; there is great debate over what is constitutive of any given gender identity. What, then, are we talking about when we discuss, e.g., the identity "woman"?

On Rea's account, gender identities are real social concepts whose content is derived from communal usage. He allows that different people, however, can hold different versions of these concepts and require different conditions for identification with these concepts. Let's say Anna uses the pronoun "she" and thinks that this is necessary and sufficient to be a woman, whereas Jess has XX chromosomes and thinks that this is necessary and sufficient to be a woman. Jess and Anna both represent themselves as women, but they clearly have radically different concepts of "woman"; moreover, it's highly possible that Jess thinks that Anna is not really a woman and vice versa. Consider how frequently cis women argue that trans women are not "real" women because a trans woman's definition of "woman" contradicts a cis woman's definition of "woman." It is not clear how Rea intends to reconcile his claim that the content of gender concepts arises from communal usage with his concession that different people may assign contradictory definitions to gender. Since the content of a gender concept varies widely from person to person, it seems wrong to say that this content comes from communal usage. Therefore, Rea's argument that gender identities are real attributes fails. But though Rea's argument fails, is there another way to argue that gender is a real attribute? It seems clear that for an attribute to be real, there must be some mind-independent content of that attribute that remains fixed regardless of to whom the attribute applies. This is not the case for gender; therefore, gender is not a real attribute.

There is, however, a different aspect of Rea's view that will be useful to defend (at least in part). One of my aims in this paper is to identify what is going on when someone makes a personal gender statement. To this end, Rea's argument that a person has privileged epistemic access to her

internal experiences seems correct. As we determined in §1, an externalist account of gender as a social class fails because it does not capture transgender individuals. Therefore, there must be some internal aspect to gender that causes some individuals to reject externally conferred gender norms. Moreover, it is intuitively clear that a person has more insight into their own internal life than others do. This is what drives Rea's intuition that one's self-conferred gender is one's "true" gender — on his view, regardless of what others perceive my gender to be, the gender I confer on myself is authoritative because I have more knowledge of my internal experience than anyone else. But as we've seen, Rea's idea of one's "true" gender as a real self-conferred attribute is mistaken.

If gender is not a real attribute, what is it? So far, I have determined that sex assignment confers a gender concept on a person, that gender is neither an externally conferred social class nor a self-conferred real attribute, and that gender arises from some kind of internal state of being. Using these premises, I will now argue that gender ought to be understood as a person's internal relationship to the gender concept conferred on them by their sex assignment, and that a personal gender statement such as "I am a woman" is an expression of this internal relationship.

§3. Gender as a conceptual relationship

At birth, an individual is assigned one of two sex categories based on external genitalia and presumed internal reproductive anatomy. Immediately, a concept of gender, and the set of norms associated with it, is conferred on the individual based on their assigned sex (Ásta 2011). The content of this gender concept and its associated norms are dependent on social and familial context. The gender concept conferred on a girl born in Tehran will likely differ from the concept conferred on a girl born in New York. Even within a social context, the content of gender might differ: the concept of gender conferred on a girl born to a conservative family in New York will

differ from the concept conferred on a girl born to a liberal family in New York. The latter girl might be raised to think gender is based on self-identification rather than biological features; the former might be raised to think gender is based on genitalia. Note that what I refer to as a "gender concept" is not the same as Haslanger's "social class" notion of gender. Here, the "gender concept" is *what one is told* is necessary and sufficient to be that gender in a certain context. While this gender concept always entails gender norms (one must strive to meet the conditions for their assigned gender concept or risk alienation), it is not a political class by nature. Moreover, Haslanger's account focuses on how one is perceived or "marked" by others for privilege or subordination in certain contexts. By contrast, the crux of my account is an individual's internal relationship to their assigned gender concept and how they express that relationship to others.

When I say that "a concept is conferred" on an individual, I mean that the individual is essentially instructed to form an idea of themselves based on a concept (Ásta 2011). A female-assigned child is told she is a girl and forms her conception of herself based on interaction with the concept "girl" (and "woman" as she transitions into adulthood). This interaction is multifaceted and involves actualization or rejection of various elements of the concept and various norms associated with the concept. A female-assigned person might hold a concept of "woman" that defines a woman as someone who has a vagina, uterus, ovaries and XX chromosomes, and who ought to have children and have a generally caring and gentle personality. This person might reject the norm of child-bearing but accept the norm of having a gentle personality; as a result, she will strive, possibly unconsciously, to have a gentle personality and thus actualize an aspect of the "woman" concept. She might reject the idea that having XX chromosomes is necessary to be a woman, but accept that having a vagina is necessary to be a woman. The sum of these interactions (i.e., actualization or rejection of various aspects of the gender concept) is a person's relationship

with the gender concept that is assigned to them. This relationship with the assigned gender concept is what constitutes a person's gender.

It is probable that virtually no one accepts and actualizes every aspect of, and norm associated with, the concept of gender that is assigned to them. Each individual likely has a slightly different relationship with their gender concept. In that case, are there any gender categories to speak of, or does each person have a unique gender? I believe the latter is the case. I said at the beginning that my view is a critical descriptive one; here arises the critical part. Like Haslanger, I am developing a view with an eye toward eliminating gender- and sex-based injustice. To that end, it's a perfectly acceptable consequence of my view that each person has their own unique gender, that is to say, that there are no gender categories at all. As it stands now, gender concepts prescribe hierarchical social roles, economic status and sexual relationships based on nothing more than a misleading dimorphic notion of human sex (Fausto-Sterling 1993). To show that these concepts are not real attributes of people, but instead concepts that are conferred onto people and with which we then form a relationship, is to delegitimize the hierarchies that they perpetuate.

I have sketched a somewhat anti-intuitive account of what gender is. On my view, each person has their own unique relationship to their assigned gender concept. Then, since gender is nothing more than one's relationship to their assigned gender concept, each person has their own unique gender, rendering gender useless as a category or identity. What, then, is going on when someone makes a personal gender statement, e.g., "I am a woman"? Since gender is not a real attribute or identity, they are clearly not conferring an identity upon themselves. I believe these personal gender statements express the speaker's relationship to their assigned gender concept. When a person whose assigned gender concept is "woman" says "I am a woman," they are expressing overall acceptance and actualization of their assigned gender concept (even if they

reject some aspects of it). When a person whose assigned gender concept is "man" says "I am a woman," they are expressing overall rejection of their assigned gender concept. When a person says "I am non-binary," they are expressing overall rejection of their assigned gender concept, regardless of what that concept is.[7]

§4. Avoiding mere dispositionalism

It may be objected that the account of gender I've provided is simply a variation on the dispositional account of gender, which posits that one is a woman if she is disposed to call herself a woman, and likewise for any other gender. Some might think that the conceptual relationship I've described is just a kind of psychological disposition. There are, however, important features of my account that distinguish it from dispositionalism; moreover, my account avoids certain shortcomings of the dispositional view. Whereas the dispositional account concerns an individual's psychological state of being, my conceptual relationship account concerns an individual's active relationship with their assigned gender concept. The latter may involve psychological elements, but will also be affected by political beliefs, upbringing, education, and other external elements and life events. Dispositionalism fails to take into account the effect of living in a gendered world. If we were not sexed and assigned gender concepts at birth, we would have no psychological disposition to identify as any gender.

Of course, if we were not assigned gender concepts, we would have no conceptual relationship with them! But if our aim is to describe what's going on now, in our current gendered

[7] Often, but not always, someone who says "I am non-binary" is also expressing rejection of the notion of gender concepts as a whole. As I touch on when discussing neo-genders in §4, personal gender statements can communicate more than the core relationship to a gender concept (e.g., political beliefs), but they always communicate the core relationship at minimum.

world, to describe gender as mere psychological disposition seems insufficient. Any thorough account of gender must include the central role played by the gender concepts conferred on us.

Finally, a significant shortcoming of the dispositional view is that it fails to capture individuals who use neopronouns, such as xe/xir, or neogenders, such as demigender. It's hard to imagine how one could be psychologically disposed to identify as any one in particular of the myriad of neogender identities. On the conceptual relationship view, however, we can say that such individuals are expressing rejection of their assigned gender concepts — indeed, a complete rejection of the notion of gender concepts as a whole.

§5. Consequences and considerations

There are a few interesting consequences of this conceptual relationship view. First, it should be clear that a personal gender statement might not always be interpreted correctly. Consider a trans man whose assigned gender concept based on his sex is "woman," but who passes visually as a man. When he says "I am a man," anyone without intimate knowledge of his sexual assignment will assume that he is expressing acceptance and actualization of his assigned gender concept. This is not, however, an issue for my view, because it's common for meaningful statements to be misinterpreted, and such a misinterpretation does not undermine the meaning of the statement. In the same vein, on this view it cannot be said that transgender individuals are "mistaken" about their genders when they make personal gender statements. If a personal gender statement is understood to be an expression of one's relationship with their assigned gender concept, then a personal gender statement can't be mistaken or untrue. There is no mind-independent fact of the matter about one's gender.

This view also might make it easier to make sense not only of neogenders, as discussed in §4, but also of absurdist gender terms such as "froggender," "fairygender," etc. We are tempted to think that people who express these neogenders are simply using nonsensical terms or even acting in bad faith. On the conceptual relationship view, however, identification with an absurdist gender ought to be understood as a complete rejection of gender concepts — a statement that we are moving past the need for gender concepts to structure our social relationships.

Another somewhat recent cultural phenomenon is the presentation of personal pronouns. Many people, some of whom might not even consider themselves transgender, use pronouns that don't seem to "match" their professed gender. For example, author and activist Leslie Feinberg, who identified as a lesbian, preferred the pronoun "he" in "all-trans settings," but preferred the pronouns "she" or "ze" when among non-trans people (Tyroler 2014). The use of the pronoun "he" by a self-identified lesbian might appear contradictory or confusing, but the conceptual relationship view of gender helps us to understand Feinberg's intentions. His use of the pronoun "he" expressed rejection of certain aspects of the female gender concept conferred on him. But his identification as a lesbian and as female-bodied expressed acceptance of other aspects of the same concept. Through various methods of expression (e.g., pronouns, self-identification as a lesbian, as female-bodied, as butch, and as transgender) Feinberg, on my view, was communicating his relationship with the gender concept conferred on him. The apparent lack of cohesion between his expressions is resolved when we view his expressions as collectively reflecting his internal conceptual relationship.

Finally, it's worth considering how this view applies to transgender individuals who undergo sex reassignment surgery. The set of norms associated with a gender concept includes norms governing both reproductive anatomy and general anatomy. Let's say Charlie was born with

female anatomy and was therefore assigned the gender concept "woman," a majority of the content of which he rejects. But after Charlie undergoes sex reassignment surgery, he has sexual anatomy that is categorized as male. In virtue of his new "male" anatomy, Charlie is now assigned the gender concept "man." Therefore, he now has an accepting and actualizing relationship with the concept "man" rather than a relationship of rejection with the concept "woman." This narrative seems to fit with the testimony of many trans people that sex reassignment surgery relieves the mental discomfort, i.e., dysphoria, that arises from rejecting anatomical norms.

§6. Conclusion

Having rejected both an externalist account of gender as a social class and an internalist account of gender as a real self-conferred identity, I have argued that gender is constituted of an individual's relationship to the gender concept that is assigned to them in virtue of their sex. Furthermore, I have argued that a personal gender statement such as "I am a woman" ought to be understood not as conferring an identity on oneself, but instead as an expression of one's relationship to their assigned gender concept. This view addresses a number of practical considerations. It fully captures the experiences of transgender individuals, provides a metaphysical framework oriented toward the elimination of sex- and gender-based injustice, eliminates the possibility that personal gender statements can be mistaken, and facilitates understanding of the continuous evolution of gender terms.

Works Cited

Fausto-Sterling, Anne. 1993. “The Five Sexes.” *The Sciences* 33 (2): 20–24. https://doi.org/10.1002/j.2326-1951.1993.tb03081.x.

Haslanger, Sally. 2000. “Gender and Race: (What) Are They? (What) Do We Want Them to Be?” *Noûs* 34 (1): 31–55. https://doi.org/10.1111/0029-4624.00201.

Rea, Michael. 2022. “Gender as a Self-Conferred Identity.” *Feminist Philosophy Quarterly* 8 (2).

Sveinsdóttir, Ásta Kristjana. 2011. “The Metaphysics of Sex and Gender.” In *Feminist Metaphysics*, edited by Charlotte Witt, 47–65. Springer.

Tyroler, Jamie. 2014. “Transmissions - Interview with Leslie Feinberg.” Blog. Camp. November 23, 2014. https://web.archive.org/web/20141123060911/http://www.campkc.com/campkc-content.php?Page_ID=225.

Rethinking the Exclusionary Rule: Rights vs. Deterrence Rationale

Anika Jain — *University of North Carolina at Chapel Hill*

Introduction

In *Weeks v. United States*, the Supreme Court adopted the exclusionary rule. The rule states that evidence obtained as a result of illegal search or seizure in violation of the Fourth Amendment cannot be used in a criminal trial.

When it was created, the rule's purpose was to protect the fundamental rights of an individual and protect judicial integrity. The justification of the rule was not to prevent police officer misconduct; rather it intended to protect 4th amendment rights through judicial review. However, over time, the justification for the exclusionary rule has evolved. This evolution is highlighted in four Supreme Court Cases: *Weeks v. United States, Mapp v. Ohio, United States v. Leon,* and *Hudson v. Michigan.* Today, the original justification for the exclusionary rule (ER) has been replaced with a new justification: to deter police officer misconduct. However, reliance on the exclusionary rule has been criticized due to its ineffectiveness in reaching this goal.

While multiple tort remedies have been proposed, such as torts, injunctions, and criminal sanctions as effective alternatives to deterring police officer misconduct. This paper argues for the imposition of both the tort remedy and the exclusionary rule for different reasons: the ER (under a rights-protection rationale, rather than a deterrence rationale) should be imposed to protect defendants' 4th Amendment rights, and the tort remedy (with a relaxing of the qualified immunity doctrine and respondeat superior liability for police departments) should be imposed both for victim-compensation reasons and for deterrence-of-police-misconduct reasons.

This paper will start by critiquing the exclusionary rule by demonstrating how it is not intended nor suited for the task of deterring police misconduct. I will then move into the successes and failures of the tort remedy, as it was before *Mapp*. Finally, I will describe whether and how these failures can be overcome.

Rights-Protection Rationale vs Deterrence Rationale

Since adopting the exclusionary rule in *Weeks,* the Supreme Court has justified its use through the rights rationale and the deterrence rationale. Originally, the Supreme Court utilized the individual rights rationale in which the goal of the exclusionary rule was to protect fundamental rights through judicial review. In *Mapp*, the court emphasizes that the exclusion doctrine as an essential ingredient in the right to be free from illegal search and seizure. Thus, the court recognizes that the use of unconstitutionally obtained evidence would in itself be a violation of 4th amendment rights. As such, unconstitutionally obtained evidence must be excluded because all defendants' fundamental rights ought to be protected.

In *Leon* and future court cases, the court justifies the use of the exclusionary rule through the deterrence rationale. This rationale is that the exclusion of evidence protects the fundamental right to not be illegally searched. Through the exclusion of evidence, the court hopes to disincentivize police misconduct by removing the incentive to disregard 4th amendment rights. In accordance with the deterrence rationale, and contradictory to the individual rights rationale, the court states in *Leon* that excluding illegally obtained evidence is not a constitutional right, therefore, in instances where the social cost of the exclusionary rule is greater than the deterrence effect it need not be applied.

Often these rationales work together. By deterring the violation of illegal search and seizure, the court protects the fundamental rights protected by the 4th amendment. Additionally, in cases where evidence is excluded the court effectively protects an individual's 4th amendment rights to not have evidence obtained unconstitutionally used against them in court.

Although these rationales can sometimes act together, these theories are distinct. The adoption of one theory over another can result in different case rulings, particularly those in which evidence is admitted because there is no misconduct to be deterred. When courts use the individual rights rationale when admitting evidence, their analysis of whether or not to apply the exclusion doctrine relies on the question of whether the evidence was obtained as a result of a 4th amendment violation.

However, as demonstrated in *Leon*, the use of the deterrence rationale requires courts to answer a different question—if the deterrence effect of the exclusion will outweigh the social cost of the exclusionary rule. This requires courts to do a cost-benefit analysis in making their decision regarding evidence exclusion. This rationale fails to recognize not having illegal evidence used against one in court as a fundamental right and instead focuses on protecting violations of illegal searches and seizures.

Further sections will analyze the costs of using the exclusionary rule under the deterrence rationale as opposed to the rights protection rationale. This new justification for the exclusionary rule requires that the Courts use a cost-benefit analysis to determine when it should be applied. I will then discuss how this cost-benefit analysis produces inconsistent case rulings and jeopardizes judicial integrity. Then this paper will look at current tort remedies and how they can effectively be used in place of the exclusionary rule in the deterrence of police officer misconduct.

Why ER is ineffective (under deterrence rationale)

The change in justification from protecting fundamental rights to deterring police officer misconduct uses the exclusionary rule as a tool for something that it was not intended for and is not effective at doing.

One of the main criticisms of the exclusionary rule is the creation of false negatives. The exclusion of probative evidence that may be crucial to convicting a defendant, under the exclusionary rule, could be inadmissible in court. Thus, defendants who are in fact guilty go free. Using the original justification for the exclusionary rule, these false negatives seem to outweigh the cost of not protecting defendants' rights. By admitting illegally obtained evidence to avoid false negatives, the court would be saying that the defendants do not have a fundamental right to not have illegally obtained evidence used against them.

In *United States v. Leon*, the Supreme Court accepts this claim in its justification for creating the good faith exception. They stated that the implementation of the exclusionary rule is based on whether the social costs of a false negative outweigh the deterrence. Thus, the purpose of the rule becomes to deter police officer misconduct. Using this new deterrence rationale justification, to create some number of false negatives does not outweigh the benefits of deterrence. This is because police officers are unlikely to be deterred by whether evidence is included or excluded. By using this new justification, the Supreme Court is allowing the guilty to walk free with no benefit.

In *United States v. Leon*, the Supreme Court acknowledged that the exclusionary rule is unlikely to have a deterrent effect on judges and magistrates because they are neutral and have no stake in the outcome of criminal prosecutions. This reasoning is also applicable to police officers. The exclusion of evidence inflicts no personal cost on the officer.

The deterrent effect on police officers relies on the hope that police officers have a personal interest in seeing someone they believe is guilty be convicted. However, this assumes that police officers have some innate sense of justice that would be offended if a guilty person were to go free. But this is unlikely to be true. The deterrent effect is meant to act on bad-faith officers. These officers are less likely than good-faith officers to have the same sense of justice the deterrence effect assumes police officers will have. Motives of bad-faith officers are likely to be scaring potential suspects and taking liberties with individual freedoms, thus, in order to deter these acts, measures that directly penalize police officers are required. Officers with malicious intent are unlikely to be deterred even if the defendant goes free. Other bad-faith officers, who knowingly commit an illegal action to get a conviction in the interest of promoting a sense of justice are also unlikely to be deterred through the exclusionary rule.

Furthermore, courts are biased toward accepting questionable police testimony in order to prevent false negatives, it seems more likely that the deterrent effect will push them to hide their illegal actions through perjury than to be deterred enough to stop. Furthermore, in many criminal justice systems, police officers are not made aware of what evidence is excluded or not. Thus, officers cannot be deterred if they are unaware of the effects of their actions.

In order to minimize false negatives, courts sometimes accept unreliable testimony so they do not have to deal with the question of not-admitting evidence. Because the exclusionary rule is only applied if there is a violation of 4th amendment rights, by showing that the collection of evidence did not violate any 4th amendment rights, the court can accept the evidence. If it seems that evidence is reliable and relevant for the conviction of a suspect, the possibility of setting a guilty defendant free seems to be undesirable to courts. It interferes with the sense of justice and responsibility courts feel toward victims and the public. When police officers perform illegal acts

while doing their jobs and these illegal acts result in incriminating evidence courts are likely to view the police officer's actions more favorably than the illegal actions of the guilty defendant because the police officer appears to have committed these acts in the interest of promoting justice. As such, courts have the incentive to accept questionable policy testimony, claiming that misconduct never occurred in order to admit the evidence.

If the original justification was used, then courts would not have to analyze whether the social costs of deterrence or false negatives is greater. By eliminating this analysis, judges will not have the option to accept questionable testimony to admit evidence. The only question that would need to be answered was if the 4th amendment was violated or not. Courts may still have the incentive to accept questionable testimony in regard to whether there was a 4th amendment violation. However, it is harder to prove that a violation did not occur than to prove that a violation occurred but not due to misconduct. Thus, by not using the deterrence rationale to justify the exclusionary rule, judicial integrity is preserved to a greater extent.

In more recent decisions of *United States v. Leon* and *Hudson v. Michigan*, the court has weighed the social cost of the exclusionary rule with its deterrence effect in order to determine whether the exclusionary rule should or should not be applied. However, this method leads to inconsistent justification for court rulings due to difficulties in quantifying the effect of police misconduct. The main reason for this difficulty, is that if the exclusionary rule does, in fact, deter police officers, it will produce a non-event that is unobservable. Therefore, the use of this cost-benefit analysis, for elements that cannot be quantified, is akin to the Courts simply guessing whether the deterrent effect will be more or less likely based on the case. This leads to inconsistent justifications in deciding cases. In some cases, like *United States v. Leon* and *United States v. Calandra*, the court has stated that the rule's inability to deter police misconduct is justification

for including evidence. In others like *Hudson v. Michigan*, the court has stated that deterrence is so great that it will result in the over-deterrence of police officer misconduct. This is a slippery slope, because whenever the Court wants to admit evidence, they can reason that the deterrence effect is not greater than the social cost. Because deterrence cannot be measured, judges must make their decision solely based on their perceived social cost of using the exclusionary rule. If the original justification was used, this incomplete cost-benefit analysis would not need to be used. Rather, the court would have to interpret whether there had been an improper exercise of power by another branch of government. This will reduce inconsistent decisions and strengthen the court's legitimacy.

Furthermore, the task of judges should be deciding the case before them without the responsibility of influencing or being influenced by other branches of government. It should not be to attempt to deter police officer misconduct—an attempt that will most likely be futile. In *Mapp*, the Court emphasizes the value of judicial integrity in protecting the rights of the parties experiencing litigation. Judges could apply the remedy depending on the context of cases being litigated without being dependent on the actions of other government branches. With the deterrence rationale, judges are burdened with the additional task of determining how the inclusion of evidence will affect the actions of police officers–a third party not involved in the case. Without any convincing data that suggest the deterrence effect of the rule, it is impossible for judges to do anything more than guess about the likelihood of deterrence. As such, the decisions they make are likely to be uninformed. Using the original justification, judges will be able to utilize their judicial discretion to analyze the facts of the case and interpretation of the 4th amendment to make an informed decision about the case rather than doing guesswork. This will preserve judicial integrity by utilizing a judge's judicial expertise to decide a case.

Supreme Court on Alternative Remedies

The Supreme Court mentioned the efficacy of alternative remedies in comparison to the exclusionary rule in *Hudson v. Michigan*. The court had forgone the use of the exclusionary rule stating that the deterrence of officer misconduct provided by civil-rights violations and internal police discipline had strong deterrent effects, perhaps even stronger than the exclusionary rule. Thus, they reasoned that use of the exclusionary rule would cause overdeterrence.

The dissent pushed back on this, arguing that the exclusionary rule remained the more effective deterrence. They claimed that the alternative remedies for police deterrence that were present at the time of *Mapp* were deemed "worthless and futile" (Scalia 2005, 609). According to the dissent, *Mapp*'s implementation of the federal exclusionary rule solved the problem of inadequate state remedies in achieving this goal. Thus, consistent with the reasoning during *Mapp*, the dissent argues that the alternative remedies for police misconduct in states during the time of *Hudson v. Michigan* are also inadequate, thus requiring the use of the exclusionary rule as a deterrent. However, this paper argues that alternatives to the exclusionary rule, in particular, tort remedies, are undoubtedly more effective in deterring police misconduct than the exclusionary rule. Implementation of this remedy will allow for the deterrence of police officer burden to be lifted off of the exclusionary rule.

It is only reasonable to assume that police officers will be deterred from committing an action when they are 1) aware that this action will result in a certain consequence and 2) when the consequence directly affects the police officer and outweighs the benefit of committing a constitutional violation. Of course, no proposed remedy will be able to completely deter office misconduct and will certainly not eliminate all constitutional violations. A law enforcement officer, acting in good faith, can violate rights while believing he was acting constitutionally.

However, the minimization of these violations done in bad faith will result in increased social utility and is likely to improve public relations between the public and police officers. This section will begin with a discussion of the critiques of the tort remedy and then explain if and how these challenges can be overcome.

Defects in the Present Remedy

Currently, victims of illegal search and seizure have two main civil remedies: Bivens suits (against federal officers) and 1983 suits (against state officers). However, these remedies are hardly effective. In most jurisdictions, officers that are serving a warrant, even if the warrant is illegal, have a complete defense. Under these remedies, this means that there is no recovery. In cases where a warrant is not administered, officers can rely on a defense of good faith and probable cause. Thus, it is unlikely that officers are held responsible for illegal searches and seizures.

Additionally, even if the officers are found liable it is unlikely that the plaintiff will receive damages. Compensatory damages usually require that the plaintiff suffer an injury to property, feelings, or reputation. Punitive damages usually require that the officer was acting with malicious intent. Because most illegal acts done by officers are not with malicious intent and proving hurt feelings and reputation is difficult, victims of unconstitutional actions are deterred from filing suits unless significant property damages have been suffered. As a result, these remedies have little deterrent effect on police officers.

One of the major criticisms of the tort remedy is the inability of victims of violations to bring court cases against suitable defendants. The qualified immunity doctrine prohibits police officers or other government officials from being held personally liable for constitutional violations unless they violate a clearly established law. Under this doctrine, civil rights plaintiffs

have to show the defendant not only violated a clear legal rule but that there is also a prior case with functionally identical facts. This greatly reduces the number of civil rights cases that are taken to trial. Although the doctrine was established to balance the need to hold public officials accountable when they exercise power irresponsibly and the need to shield officials from harassment, distraction, and liability when they perform their duties reasonably, it has clearly become a way for police officers to dodge accountability. But if it were abolished, it is likely that police officers would be harassed with frivolous lawsuits, creating a dilemma. Either deterrence of violations does not occur or there is deterrence of people entering the law enforcement profession.

However, even if qualified immunity was eliminated, there remains a question of whether courts should require the officer to pay monetary damages. The threat of large judgments is likely to deter qualified people from becoming police officers as well as unjustly punish officers and their families for errors in judgment.

Additionally, civil prosecutions of police officers are particularly difficult. Police tend to have higher respectability which makes the jury biased in their favor. Furthermore, plaintiffs in cases where the illegal search did yield incriminating evidence are unlikely to garner juror sympathy. They are also unlikely to have the resources to take legal action from behind bars. The resistance of jurors to believe allegations of misconduct from police officers makes prosecutors hesitant to bring cases against them. Prosecutors may also choose to not bring cases against police officers because they typically work together to prosecute other alleged criminals resulting in a close relationship. In some jurisdictions, district attorneys are elected and rely on support from police unions and their supporters.

In order for a deterrent effect to occur, the remedy must encourage those whose rights have been violated to seek remedy. However, due to the difficulty in finding an attorney who is willing to bring a case against a police officer and the unlikeliness of receiving any material monetary remedy, plaintiffs are discouraged from suing law enforcement officers. For the suits that are brought, the officers experience no deterrent effect. Statistics about internal department actions show that officers who are sued are more than twice as likely to be promoted than punished. It would seem reasonable to assume police officers would reward aggressive police actions and be unlikely to punish fellow officers who discover incriminating evidence against a suspect. As such, remedies must be made to make the tort remedy a more effective deterrent.

Remedy

Before *Mapp*, eighteenth-century common law allowed suits against officers personally, but it was understood that the true party of interest was the government itself. The government would be forced to indemnify officials carrying out government policy in order to prevent deterrence from government positions. A similar remedy today would be to recognize the direct liability of the government entity. Justice Burger suggests a remedy similar to the doctrine of respondeat superior in his dissent in *Bivens*. He explains that:

"The venerable doctrine of respondeat superior in our tort law provides an entirely appropriate conceptual basis for this remedy. If, for example, a security guard privately employed by a department store commits an assault or other tort on a customer such as improper search, the victim has a simple and obvious remedy—an action for money damages against the guard's employer, the department store." (Brennan 1970, 411)

The police department is likely to seek indemnification, dock pay, require training, or otherwise discipline officers who trigger the government's liability, thus creating a deterrent effect. Additionally, seeking redress against the police department is superior to seeking redress from an individual officer. Individual officers most likely do not have the means to offer more than a minimal collectable amount. This is important because if the compensation is not adequate plaintiffs are discouraged from going through the trouble of a lengthy lawsuit.

Justice Burgers' proposal is five parts that can function as an effective deterrent for police misconduct:

1) A waiver for sovereign immunity for the illegal acts of law enforcement officials committed in the performance of assigned duties.
2) The creation of a cause of action for those individuals whose constitutional rights were violated by government agents.
3) The creation of a tribunal, quasi-judicial in nature or perhaps patterned after the United States Court of Claims to adjudicate all claims under the statute.
4) A provision directing that a civil damage remedy is completely in lieu of the exclusion of evidence obtained in violation of the Fourth Amendment.
5) A provision commanding the courts not to exclude any evidence that would otherwise be admissible but for a Fourth Amendment violation. (Brennan 1970, 411)

Burger intends this proposal to be in place of the exclusionary rule on the basis that the exclusionary rule is completely ineffective in deterring police misconduct. While it is true that the exclusionary rule is not effective in that sense, it is still incredibly important in maintaining individuals' 4th amendment rights. For this reason, the 4th and 5th elements of the proposal should not be implemented.

However, elements 1, 2, and 3 adequately address the shortcomings of the present tort remedy. The waiver of qualified immunity in cases of illegal acts would guarantee liable defendants against whom cases could be brought while also curbing the over-deterrence of the profession. The tribunal board according to Justice Burger "is likely to eliminate the problem of

jury bias" (Brennan 1970, 411). As stated by Burger, "I doubt that lawyers serving on such a tribunal would be swayed either by undue sympathy for officers or by the prejudice against 'criminals' that has sometimes moved lay jurors to deny claims" (Brennan 1970, 411). This will lead to plaintiffs being more encouraged to seek remedy after a violation of rights. This will serve as a deterrence of officer misconduct because they will fear prosecution for unconstitutional acts.

This remedy leaves concern as to how victims in cases where incriminating evidence is discovered would be compensated. It would be odd to charge a defendant with a crime and then offer a tort remedy to compensate them for the violation of their rights. This raises questions of how the compensation would be determined and for what exactly the defendant is being compensated for.

However, if this tort remedy is used in conjunction with the individual rights rationale of the exclusionary rule then the number of cases where a guilty defendant must be compensated will decrease. The implementation of the exclusionary rule using the individual rights rationale would result in more acquittals: without good faith exceptions, many court cases where evidence is not suppressed under the deterrence rationale would be suppressed under the individual rights rationale. Thus, the defendants who suffered these violations would not have been charged with the crime. The evidence that is suppressed would not be recognized by the court. Utilizing the same justification that the use of illegally obtained evidence violates a fundamental right, remedies sought after the violation must be unrelated to whether or not the evidence would have resulted in a conviction if it was obtained constitutionally. Thus, damages will be awarded assuming the victim was innocent.

One major criticism is that this approach will result in defendants being wrongly acquitted and also claiming damages. However, this negative must be accepted in order to maintain

consistent justifications in court rulings. It should not be the case that sometimes the evidence is admitted and sometimes not based on the court's guess of whether a deterrent effect will be achieved or not. Additionally, it should not be the case that police officers can commit illegal acts and not be required to compensate their victims due to the possibility that if the evidence was admitted the defendant would have likely been found guilty.

If it is the case that a defendant is found guilty due to other evidence that is not the result of a constitutional violation, then this remedy can be sought for any property damages. This is necessary to deter police officers from committing illegal acts. However, damages for injury to feelings or reputation cannot be awarded because these damages would have occurred as a result of the legal collection of evidence anyway.

Ideally, preventing these violations from occurring would be better than utilizing civil remedies to compensate victims afterward. However, the exclusionary rule does not achieve this deterrence because bad-faith police officers are not being directly affected. Using this remedy there will be a stronger deterrent effect because officers will be directly affected. This will produce a stronger deterrent effect leading to fewer instance of police officer misconduct.

Under this remedy violations in good faith will also have to be compensated by police departments. This negative must be accepted because any method of compensation that differentiates between good and bad faith cops will promote perjury as police officers who are bad faith will try to portray themselves as good faith. For negatives that result in minor injuries, civil remedies should be expected to compensate victims because illegal actions by officers of law regardless of if they caused serious or minor injury because law enforcement should abide by the laws, and not doing so insults the integrity of law enforcement.

Conclusion

The Supreme Court decisions in *Weeks*, *Mapp*, *Leon*, and *Hudson* show how the justification of the exclusionary rule has changed from protecting individuals' rights to deterring police officer misconduct. The deterrence rationale for the exclusionary rule does not justify the social costs of the rule. However, the original justification for the exclusionary rule does. This paper advocates for the justification of the exclusionary rule to be shifted back to protecting the fundamental rights protected by the fourth amendment as it will strengthen judicial integrity and the court's legitimacy. To meet the Supreme Court's goal of deterring the police alternative, tort remedies are more effective. Although it is undoubtedly more effective than the exclusionary rule, there are some valid critiques in its implementation. However, these critiques can be remedied through Justice Burgers's proposal. Thus, it becomes clear that the best approach is to reinstate the original justification for the exclusionary rule and implement a tort remedy for the purposes of police officer misconduct deterrence.

Works Cited

Alschuler, Albert W. "Exclusionary Rule and Causation: Hudson v. Michigan and Its Ancestors, The." *IOWA LAW REVIEW*, n.d., 79.

Justia Law. "Alternatives to the Exclusionary Rule." Accessed December 7, 2022. https://law.justia.com/constitution/us/amendment-04/32-alternatives-to-the-exclusionary-rule.html.

Brennan, William J., Jr, and Supreme Court Of The United States. *U.S. Reports: Bivens v. Six Unknown Fed. Narcotics Agents, 403 U.S. 388*. 1970. Periodical. https://www.loc.gov/item/usrep403388/.

Clark, Tom Campbell, and Supreme Court Of The United States. *U.S. Reports: Mapp v. Ohio, 367 U.S. 643*. 1960. Periodical. https://www.loc.gov/item/usrep367643/.

Cloud, Morgan. "Judicial Review and the Exclusionary Rule." *Pepperdine Law Review* 26, no. 4 (May 15, 1999). https://digitalcommons.pepperdine.edu/plr/vol26/iss4/4.

Collins, Allyson. "Shielded from Justice: Police Brutality and Accountability in the United States | Office of Justice Programs." Accessed December 7, 2022. https://www.ojp.gov/ncjrs/virtual-library/abstracts/shielded-justice-police-brutality-and-accountability-united-states.

"C-SPAN Landmark Cases | Season One - Home." Accessed December 7, 2022. https://landmarkcases.c-span.org/Case/9/Mapp-v.-Ohio.

Day, William Rufus, and Supreme Court Of The United States. *U.S. Reports: Weeks v. United States, 232 U.S. 383*. 1913. Periodical. https://www.loc.gov/item/usrep232383/.

"Exclusionary Rules—Is It Time for Change? | SpringerLink." Accessed December 7, 2022. https://link.springer.com/chapter/10.1007/978-3-030-12520-2_12.

Frankfurter, Felix, and Supreme Court Of The United States. *U.S. Reports: Wolf v. Colorado, 338 U.S. 25*. 1948. Periodical. https://www.loc.gov/item/usrep338025/.

Hilton, Alicia M. "Alternatives to the Exclusionary Rule after Hudson v. Michigan: Preventing and Remedying Police Misconduct." *Villanova Law Review* 53 (2008): 37.

Jr, Myron W Orfield. "The Exclusionary Rule and Deterrence: An Empirical Study of Chicago Narcotics Officers." *The University of Chicago Law Review*, n.d., 54.

"Narrowing Application of the Exclusionary Rule :: Fourth Amendment -- Search and Seizure :: US Constitution Annotated :: Justia." Accessed December 7, 2022. https://law.justia.com/constitution/us/amendment-04/35-narrowing-application-of-the-exclusionary-rule.html.

Redemann, Bob. “The Historical and Philosophical Foundations of the Exclusionary Rule.” *TULSA LAW JOURNAL* 12 (2013): 15.

Scalia, Antonin G, and Supreme Court Of The United States. *U.S. Reports: Hudson v. Michigan, 547 U.S. 586*. 2005. Periodical. https://www.loc.gov/item/usrep547586/.

“The Tort Alternative to the Exclusionary Rule in Search and Seizure.” *The Journal of Criminal Law, Criminology, and Police Science* 63, no. 2 (1972): 256–66. https://doi.org/10.2307/1142302.

Wilson, Jerry V., and Geoffrey M. Alprin. “Controlling Police Conduct: Alternatives to the Exclusionary Rule.” *Law and Contemporary Problems* 36, no. 4 (1971): 488. https://doi.org/10.2307/1190931.

Counting Coincidences: A Response to Fine's Counterexamples Against Locke's Thesis

Irene Wang — *University of Michigan*

Introduction

Locke's Thesis states no two material objects that belong to the same sortal can exist at the same place at the same time. This denies the possibility of spatial-temporal coincidence of material objects. This thesis speaks to our intuition that the exact region of space occupied by a particular object cannot be simultaneously occupied by another distinct object. But intuition alone is not enough to support this thesis. In Fine's "A Counter-Example to Locke's Thesis", three counterexamples concerning coinciding letters were raised against Locke's Thesis (Fine, 2000). I argue that Fine's counterexamples fail to present a genuine challenge to Locke's Thesis. Locke's Thesis only applies to material objects belonging to sortals that have a discrete counting criteria with specific, well-defined individuation conditions. Individuation conditions are the set of criteria by which we distinguish an object belonging to a particular sortal from everything else in the universe (Oderberg, 1996, p. 147). The sortal "letter", as presented by Fine, has ambiguous individuation conditions that are open to interpretation. The problem of coincidence raised by Fine is due to the lack of a specific and stable counting criteria associated with letters. Thus, "letter" is not a proper material object sortal to be considered under Locke's Thesis. In this paper, I will provide a brief introduction to sortals and propose the Sortal Counting Criterion to evaluate whether something is a valid sortal applicable by Locke's Thesis. I will examine Fine's counterexamples and offer several examples that demonstrate the problem of basing counting

criteria upon external relations. Lastly, I will explore how counting relies on conceptual boundaries in order to provide definitive answers to the "how many" question.

Getting Sortals All Sorted Out

Given the myriad of material objects that we encounter in our daily lives, it is helpful to group objects that share similar characteristics into different buckets of categorization, or "sortals". Though no unified definition of "sortal" or "kind" currently exists. According to Grandy and Freund (2021), a sortal must fulfill three crucial criteria:

1. A sortal tells us the essence of an object.
2. A sortal tells us how to count objects that belong to it (i.e. individuation conditions and counting criteria).
3. A sortal tells us when an object continues to exist and when an object goes out of existence (i.e. persistence conditions for the object).

In this paper I will focus on criterion 2 for evaluating sortals. I propose a more stringent version, which I shall term the Sortal Counting Criterion (abbreviated SCC), as follows:

> **SCC:** A valid sortal *S* must provide a set of specific and stable counting criteria *C* such that for all *x*s belonging to sortal *S*, the *x*s can be individuated and counted in an undisputed manner according to *C*.

SCC does not permit any vagueness (criteria C must be specific), nor mutability (criteria C must be stable) of counting criteria for any given sortal. I will not be examining how sortals in general are defined and chosen. I will simply be evaluating whether a given candidate is a valid sortal according to this criterion.

Cases of Coincidences? Well, Sort of…

Fine (2000) lays out the following four premises for his chosen counterexamples involving the proposed sortal "letter".

1. The letters are the same sort.
2. The letters spatially coincide (i.e. are at the same place at some time).
3. The letters are distinct.
4. The letters are "things" in the sense relevant to the application of the thesis.

Fine then gives the Bruce-Bertha Letters, Bruce-Neighbor Letters, and Prittle-Prattle Letters as counterexamples. In the Bruce-Bertha Letters example, husband Bruce writes a letter to his wife Bertha using scorch marks on one side of a single sheet of paper. Bertha then replies by writing with scorch marks on the other side of the same sheet of paper. Fine argues that the Bruce-to-Bertha letter is now spatially coinciding with the Bertha-to-Bruce letter. In the Bruce-Neighbor Letters example, Bruce writes a letter to Bertha on one side of a single sheet of paper, while his neighbor writes a letter to his wife on the other side of the same sheet of paper simultaneously. Fine argues that the Bruce-to-Bertha letter spatially-temporally coincides with the neighbor-to-wife letter. Lastly, in the Prittle-Prattle Letters example, father Fluent writes to his two daughters using the same set of symbols on a single sheet of paper. The symbols are interpreted by Fluent's elder daughter in the language Prittle, where the letter is addressed to the elder daughter only. The symbols are interpreted by Fluent's younger daughter in the language Prattle, where the letter is to the younger daughter only. Fine proposes that the Prittle Letter necessarily coincides with the Prattle Letter. Both letters share the same persistence conditions and spatial-temporally coincides during the entirety of their existence (Fine, 2000, pp. 357-360). These cases seem to be progressively challenging counterexamples to Locke's Thesis. Premises 1, 3, and 4 are based on the assumption that "letter" is a valid sortal pertinent to Locke's Thesis. I shall demonstrate that unfortunately, "letter" is not a valid sortal as it fails the SCC. I will be focusing on the Prittle-

Prattle letter example as it appears to pose the most serious threat to Locke's Thesis due to their necessary spatial-temporal coincidence.

Locking Down Locke's Thesis

Fine's (2000) version of Locke's Thesis states:

> **LT**: No two things of the same sort can be in the same place at the same time.

Locke's Thesis denies the possibility of spatial-temporal coincidence of material objects that belong to the same sortal. At first glance, this thesis seems rather vague. It is unclear what "things" exactly refers to. This crucial definition is intimately tied to Fine's defense of his fourth premise - letters are "things" that Locke's Thesis applies to. Fine acknowledges that Locke's Thesis should only be applied to proper material objects. The thesis does not apply to "objects that are dubiously material", such as shadows or clouds. It also does not apply to "objects of dubious individuality", such as electrons or bosons (Fine, 2000, p. 359). Furthermore, the object in question must belong to a valid sortal by the SCC criteria outlined above. Locke's Thesis only applies to objects belonging to valid sortals, that is, objects that are strictly material with specific and stable individuation conditions. Shadows, clouds, electrons, and bosons are not valid sortals because of their vague and imprecise individuation conditions, which makes discrete counting difficult if not impossible.

The specific object Fine chose to discuss is "letter". What makes something a letter? Conventionally, a letter entails symbols instantiated over a surface that conveys information from the sender to the recipient. By this definition, a text message could be counted as a letter. Perhaps this definition is too broad and abstract to be useful. Fine did not explicitly provide a specific definition of letter, but rather relies on the reader's intuitions and everyday conceptions of letters.

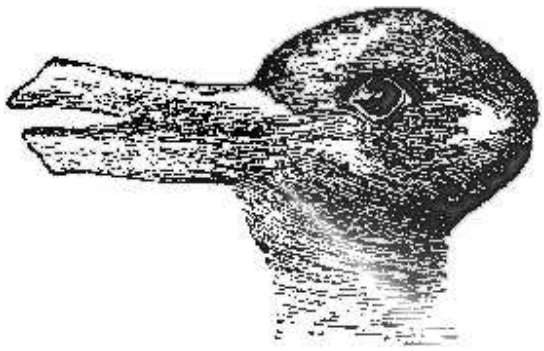
This leaves the individuation conditions and counting criteria of letters open to the issue of vagueness and mutability. Fine starts out describing letters as material objects by referring to their properties of being able to be "stacked, weighed, damaged, destroyed, and so on" (Fine, 2000, p.359). Taken *prima facie*, it seems Fine has chosen a fine example of material object. The particular description of "letter" provided by Fine refers to the pieces of paper with symbols that can be used to convey information from a sender to a receiver. The point was emphasized with the statement "once a possibly abstract sense is set aside" (Fine, 2000, p. 359). By this definition, the object(s) in question purely concerns the pieces of paper, independent of any abstract notion formed by external relations. Yet this abstract sense did not get set aside. Rather, Fine relied on the differing external relations the letter bears to its sender and receiver to make the case for coincidence. In the Prittle-Prattle letter case, the semantic content of the symbols on the paper is used to serve as the basis for individuation rather than the paper itself.

The letter to paper relation resembles that of a statue and the clay that constitutes it. The clay is a strictly material object, and so is the paper. However, it is unclear whether "statue" is a strictly material object. It seems "statue" refers to the shape of the clay, which is a property. In a similar vein, "letter" refers to the semantic content of the symbols on a piece of paper, which picks out a certain property that is contingent on the presence of interpreters who can extract such contents. The symbols can be said to be a certain feature or property of the paper, much like the color of the paper. A red piece of paper, for example, appears red to those with normal colored vision, but appears gray to those with red-color blindness. The paper itself did not "change" colors, but appears as different colors depending on the observer. Similarly, the Prittle-Prattle letter is instantiated by a single set of symbols, but appears different (in terms of semantic content) to each of Fluent's daughters. If we restrict the domain of reference to a single observer, then there would

be no issues regarding coincidence. The type of "coincidence" instantiated by Fine's counterexamples is generated via different external relations the target has with various observers, rather than the object itself. I will explore the issue this type of externally dependent "coincidence" raises with respect to the counting criteria next.

Illegal Letter and Interesting Illustration: Some Counter-Counterexamples

Suppose instead of arriving at the hands of Fluent's daughters, the Prittle-Prattle letter gets mistakenly delivered to Lingo, who not only speaks both Prittle and Prattle but is also able to take any given set of symbols and construct a multitude of artificial languages from it. Knowing that opening someone else's mail is a felony, Lingo decides to create a third language using Fluent's letter. Interpreted using this third language, the letter is now addressed to him. Voila! Crime avoided. It is not difficult to see that for any given set of symbols, an infinite number of "letters" can be instantiated by mere shifts in interpretation. Letter, as defined and used in the manner by Fine, does not provide a stable counting criteria, as the number of letters in any given region of space can proliferate indefinitely via different external relations. Thus, letter fails to be a valid sortal by the SCC.



The problem with interpretation can be further illustrated by the rabbit-duck optical illusion (Jastrow 1899, p. 312). If we ask "how many rabbits are there", we can provide a definitive answer: one. Similarly, if we ask "how many ducks are there", we can provide a definitive answer: one. In this case, the term "rabbit" and "duck" are both valid sortals since there is an unambiguous set of criteria that the image fulfills to classify it

as a single “rabbit” or a single “duck”. Now, suppose I were to ask “how many images are there”, it is unclear what the answer is supposed to be. If we define a single “image” to be the visual stimuli over a specified region of space, or perhaps a unified work of art as indicated by the artist, then the answer is going to be one. However, if we define a single “image” to correspond to a unique interpretation by the viewer, then the answer is going to be two - one for the duck interpretation and one for the rabbit interpretation. “Rabbit” and “duck” would be valid sortals with respect to the above drawing. But “image” would not, since the counting criteria for “image” is ambiguous. Another example would be the Rorschach ink blots. In this case, each inkblot is designed to be open to the viewer’s interpretation. For a given ink blot image, if we ask “how many Rorschach inkblots are there”, the answer is going to be one. If we ask “how many interpretations of the inkblot are there”, the answer is indeterminate, as each viewer can have any number of unique interpretations for a given inkblot.

Similarly, Fine construes his counting criteria of “letter” in a manner akin to the number of interpretations instantiated over a given region of space. Interpretations are inherently dependent on an object’s external relations with interpreters, and thus are not intrinsic properties of the object. If we remove such external dependencies, the only material object that is present is a single sheet of paper with written symbols on it. It is the paper that is a proper sortal and material object subject to Locke’s Thesis. We can individuate sheets of paper unambiguously given our everyday conceptions of what a single sheet of paper entails. Letters, on the other hand, is an artifact with less well-defined individuation conditions that can be exploited to construe cases of coincidence, such as the counterexamples presented by Fine. Thus given the SCC, letters fail to be a valid sortal. This does not imply that letters are not material objects, only that letters are not the proper sort of

objects that Locke's Thesis applies to. The ontological status of letters, statues and other artifacts will not be examined here.

So, What Sort of Things Count?

Okay, enough with letters. One might wonder if there is anything that could be said about establishing counting criteria more generally. First, we must establish what it is that we are counting. As mentioned earlier, the counting target in question must belong to a valid sortal that provides specific and stable individuation conditions. Thus if we have a valid sortal, we would be able to arrive at an undisputed count of objects belonging to that sortal per the sortal's individuation conditions. Elaborating on Frege's counting criteria for concepts, Koslicki proposes the individuation condition as either *discreteness* or drawing of *conceptual boundaries*. Discreteness here refers to the absence of spatial overlap between objects that are being counted (Koslicki, 1997, p. 405). This seems to be an intuitive way to separate objects by their visual boundaries. Material objects, by this account, are presented as "neatly separated parcels", with well-defined boundaries (Koslicki, 1997, p. 410). The assumption here is that our visual perceptions of object boundaries correspond with how they are ontologically. It is unclear if that is the case. Such criteria also run into the issue of vagueness with objects that lack clear visual boundaries but are nevertheless distinct (such as in the cases of shadows and clouds).

The second proposal of individuation based on conceptual boundaries is a compelling suggestion that accounts for our everyday practice of counting objects. Objects, existing as they are, do not come prepackaged into countable units. Rather the work of carving up our reality into distinct units is "done by our concepts" (Koslicki, 1997, p. 416). Concerning material objects, there is an important *mass-count* distinction. Objects with *count-occurrence* can be counted using discrete units (ex. apples, cows, lamps). Objects with *mass-occurrence* cannot be counted using

discrete units, but can only be referenced in terms of continuous quantities (ex. water, gold, sugar). Count-occurrence corresponds to the "how many" question and mass-occurrence corresponds to the "how much" question (Koslicki, 1997, p. 404). This distinction is caused by one of granularity by which we perceive these objects, not due to ontological differences between the two categories, as both occurrences involve material objects. Count-occurrence objects have more "natural" visual boundaries that distinguish them from all other objects. The noun we attribute to such objects serves as a discrete counting unit in and of itself (it makes sense to ask "how many apples", without specifying an additional unit). On the other hand, mass-occurrence objects, such as water, cannot be counted by themselves (it does not make sense to ask "how many water"). But such objects can be counted once we apply a discrete counting unit (such as "how many glasses of water", or "how many liters of water"). In these cases, the "work" of isolation is being done by the counting unit, for we do not directly perceive the object in a discrete manner.

Objects with count-occurrence are valid sortals per the SCC since they have specific, stable, and discrete counting criteria. Objects with mass-occurrence can fulfill the SCC with the aid of a discrete counting unit, but cannot be individuated by themselves. It is unclear if this need for an additional unit posits an ontological difference between the two types of objects. If we consider Koslicki's proposal of isolation as conceptual boundaries, then drawing the conceptual boundary of an apple versus drawing the conceptual boundary of a glass of water should involve the same process. Once we have decided on the specific conceptual boundaries of an object, we can then use that boundary to carve reality into individual objects. From here, we can count objects as discrete entities.

One might argue this issue of vagueness surrounding letter's counting criteria can be avoided by providing a specific definition of such criteria at the beginning of one's inquiry. If we

have a set of specific and stable individuation conditions for letters, then "letter" would be a valid sortal applicable for Locke's Thesis. Unfortunately, no such unified definition was provided by Fine. In fact, this is the very loophole that has been exploited to make the case for the counterexamples. Perhaps the lack of clear elucidation of "letter" as a sortal does not completely nullify Fine's cases if we can come up with a better definition of "letter". If we define a letter as a coherent set of semantic content instantiated by a sender to a receiver, then we run into the problem of indeterminate number of interpretations as demonstrated by the Lingo thought experiment. This can be curbed if we limit the number of interpretations to that specified by the sender. However, this definition is still contingent upon the sender's mental state (which is an external relation) and not upon the physical object itself.

If we define letters purely as the sheet of paper that constitutes it, then in all three cases, there is only a single sheet of paper. Thus Locke's Thesis is upheld. However, by this definition, any piece of paper with symbols that have the *possibility* of carrying meaning can be counted as a letter. This would mean a paper with nothing but purple polka dots, would be counted as letters, as the polka dots could "possibly" carry meaning. Taken to less extreme examples, other papers with semantic content with a sender and a recipient, such as advertisements or birthday cards, could be counted as letters. The very notion of "letter" seems to be undermined. "Letters", as dictated by our everyday conception, have fuzzy conceptual boundaries much like the rabbit-duck illusion. The external relations a letter bears to its sender and recipient is integral to its existence. However, such external relations should not be used as the basis of its counting criteria as they are vaguely defined and unstable. Letter can only be a count-occurrence object when we define it in a strictly material sense, namely, the piece of paper that constitutes it. By this account, in all three of Fine's

counterexamples, there is only one piece of paper. Coincidence does not occur and thus does not pose a genuine challenge to Locke's Thesis.

Conclusion

Fine's counterexamples involving letters fail to raise a genuine challenge to Locke's Thesis. Letters do not have specific and stable individuation conditions and thus cannot be counted in an unambiguous manner. The number of letters over a given region of space is disputable depending on which set of conditions are being chosen to serve as the counting criteria. I have explored the letter-as-paper definition, which denies coincidence. As well as the letter-as-semantic-content definition, which does not have a stable counting criteria and thus fails to be a valid sortal by the SCC. Counting criteria are constructed from conceptual boundaries that carve up our reality into separate objects. These boundaries must be specific enough with suitable discrete units in order for counting to occur. When we ask how many objects are over a given region of space, we must provide a corresponding counting criteria, or else the question is misconstrued and open to interpretation. The issue with coincidence proposed by Fine is associated with ambiguous counting criteria and not the physical object itself. Any given region of space does not have inherent individuation conditions over its contents. It is our conceptual boundaries that individuate objects and make numerical quantification possible. At the end of the day, it is our concepts of the world that count.

Works Cited

Fine, K. (2000). A counter-example to Locke's thesis. *The Monist, 83*(3), pp. 357-361. https://www.jstor.org/stable/27903691

Grandy, R. E., & Freund, M. A. (2021). *Sortals*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2021/entries/sortals/.

Jastrow, J. (1899). The mind's eye. Popular Science Monthly, 54, pp. 299-312. https://www.proquest.com/magazines/minds-eye/docview/288274780/se-2.

Koslicki, K. (1997). Isolation and Non-Arbitrary Division: Frege's Two Criteria for Counting. *Synthese, 112*(3), pp. 403-430. https://www.jstor.org/stable/20117670.

Oderberg, D. (1996). Coincidence under a Sortal. *The Philosophical Review, 105*(2), pp. 145-171. http://www.jstor.com/stable/2185716.