A Priori

The Brown Journal of Philosophy

Volume 4

Brown University
Department of Philosophy
2019

Printed by IngramSpark

a-priori@brown.edu

A Priori: The Brown Journal of Philosophy is made possible through the generosity of the Department of Philosophy at Brown University. The Journal would like to offer special thanks to Professor Paul Guyer and Katherine Scanga for their indispensible contributions.

Letter from the Editor

Philosophy has been criticized for its insular and hyper-specialized nature enough times that I won't belabor the point. Susan Haack recently decried the rise of intellectual fads and fashions from "'formal' everything" to "recurrent outbreaks of galloping Gettieritis." Despite widespread concern about the state of the discipline, there is little agreement about a solution. We, the editors of A Priori, are encouraged by the latest turn towards inclusivity and interdisciplinarity in philosophy. Sub-fields such as philosophy of mind and philosophy of science have become increasingly empirically-informed, while critical theory and philosophy of race, gender, and sex are finally starting to see the rewards of their decades-long fight for a place in the canon and the classroom. But we have also grappled with an apparent tension between interdisciplinary legibility and preserving the often jargon-heavy classical problems that make analytic philosophy so intellectually satisfying. This tension makes itself felt in every phase of the editorial process. Still, with each issue we find that not only is there room for both – or better yet, many – types of philosophy, but they complement each other. We are confident that if there is room in our small volume for the breadth of inquiry that philosophy has to offer, then there is also room in the academy.

Margot Witte, Editor-in-Chief

Table of Contents

Can a Sadistic Torturer Be Convicted of Irrationality?

Val Borba

There is one main reason why we might want to convict a sadistic torturer of irrationality, and that is a "contemporary form of the Kantian hope" that morality is intrinsically tied to our capacity to reason (Mercer n.d., 5), because we might like to think that to act immorally is to act irrationally, and that any moral action is a rational one. Bernard Williams (1981) has argued that this hope is misguided and that morality is not tied to reason in this way, so we, in fact, cannot convict a sadistic torturer of irrationality (though we might convict him of being nasty). John McDowell (1995), on the other hand, has argued that we may indeed convict a sadistic torturer of irrationality;[1] if the sadistic torturer had been properly brought up, if she were a better agent than she in fact is, then she would not have any reason to sadistically torture anyone. This means she has reasons not to sadistically torture anyone, and because she has failed to see these reasons, we may convict her of irrationality.

In this essay I will argue, with Williams, that morality and rationality are not quite tied together in the way outlined

above. I will consider Christine Korsgaard's neo-Kantian approach, however, as a bridge between Williams' and McDowell's views, and argue that, whilst we cannot convict a sadistic torturer of irrationality, we can convict him of inconsistency or, in some cases, of arationality.

Now, what is the difference between irrationality, inconsistency and arationality? For the purposes of this essay I will define them as follows. An agent who is arational does not require reasons for acting. An agent who is irrational, on the other hand, is an agent who acts in a way that is contrary to what she has reason to do. Finally, an agent who is inconsistent acts on one reason she has which is in conflict with another reason she has for acting in a different way. I will return to these distinctions later.

Before we proceed we must also distinguish between three kinds of sadistic torturers, one who enjoys sadistically torturing others and values his identity as such — Caligula for example. Another who does not endorse such an identity but is weak-willed and cannot control his urges to torture others — Dr Jekyll for example. And another who neither rejects nor endorses the practical identity of a sadistic torturer and hence whose "conduct [is] like that of a wanton", who treats each and every one of his present desires as a reason for action (Korsgaard 1996, 99) — a Satyr for example.[2] I will re-

turn to these three cases below.

In *The Sources of Normativity* Korsgaard aims to establish that we all have moral reasons to act in moral ways towards others because we have reasons to value reflective agency (that is, humanity)[3] as an end in itself. Korsgaard derives such moral reasons from an analysis of the structure of the human mind. She starts from what she takes to be the rather uncontroversial claim that the human mind is "essentially reflective" (Korsgaard 1996, 92). This means we need reasons to act, we don't just act on desires, nor are we mere battlefields of conflicting desires. We are reflective, autonomous agents, meaning that we must endorse a particular desire in order to act on it. In order to endorse a particular desire, we need reasons to choose one desire over another. And in order to choose, we must act according to a law. But this is not just any law, the law that we act on must also be up to us, because otherwise we would not feel ourselves to be autonomous, which we do.[4] This law must also, and therefore be, expressive of ourselves, that is, the law that we act on depends on who we think we are, it depends on our practical identities. So, practical identities give rise to reasons for action, and if we don't act on these reasons then we can no longer think of ourselves as having such identities.

We must also have at least some conception of our

practical identity, otherwise we wouldn't have reasons to do any one thing over another. But some of our identities are more important to us than others, and some of them are more easily shed than others. So when our identities come into conflict, i.e. when they give rise to opposite reasons, we are forced to choose which of our identities is more valuable to us. But how?[5]

All of our practical identities are contingent (that is, all but one), so one or another of them may be shed. But "what is not contingent, is that you must be governed by some conception of your practical identity", because if you are not, then "you will lose your grip on yourself as having reason to do one thing rather than another" (Korsgaard 1996, 120-1). But this springs not from a particular conception of yourself, rather, it springs from your identity "simply as a human being", that is, as a reflective agent (Korsgaard 1996, 212). Thus, to be a human being is to have moral identity. And moral identity is prior to any other practical identity you might have, because it is directly derived from your very existence, from your status as a member of humanity. Beyond this, any other practical identity you might have, as a son or a mother, a rockstar or an accountant, all of them depend on your deepest identity as a member of humanity, as a moral agent. This is because if you didn't have the prior identity 'member of humanity', you would not be able to have any of the other practical identities that

you value and endorse. You cannot conceive of yourself as a rockstar if you don't also conceive of yourself as a member of humanity, and from your identity as a member of humanity, you are also a moral agent. Essentially "all value depends on the value of humanity" (Korsgaard 1996, 121). Thus, a value that contradicts the value of humanity leads to a kind of pragmatic contradiction, an inconsistency, because in order to have any value at all, even one that contradicts the value of humanity, is to value humanity in some prior sense.

Now, let's turn to the issue that arises from the Caligula-like torturer archetype, the sadistic torturer who values the practical identity of a sadistic torturer. Is this a rational position? We might agree with Korsgaard's argument above that the value of humanity is in some sense prior or deeper than any other value we might have, because any such value derives its normativity from the fact that we are humans. But we might question whether our commitment to another value may not be stronger, for example, in the case of Caligula, whom we might describe as being more deeply committed to the value of humanity, but more strongly committed to the value of sadism. That is, what if Caligula simply cares more about sadism than he does about humanity?

Faced with this problem, Korsgaard distinguishes be-

tween two kinds of conflict that can arise from our practical identities. We can have a practical identity that is "in and of itself contradictory to the value of humanity", e.g. an assassin, or we can have a practical identity that "is not by its nature contradictory to moral value, but that leads to a conflict with it in this or that case" (Korsgaard 1996, 126). Which kind of conflict arises for Caligula? It is certainly true that his practical identity as a sadistic torturer leads to conflict with the value of humanity in some cases as, by definition, he often has reason to harm or even kill others. These actions are contradictory to the value of humanity. Beyond this, it seems that Caligula never has good reason to treat the humanity of his victims as an end in itself, since his victims are merely a means to his pleasure, and so we can conclude that the practical identity of a sadistic torturer is in and of itself contradictory to the value of humanity. It's true that Caligula might value the humanity of his friends and family, but, as Korsgaard writes, this is a "reflectively unstable" position that is likely to lead the agent to reflection. (Korsgaard 1996, 128) And when he reflects more deeply, he will go on to see his mistakes, and to have reasons to shed his identity as a sadistic torturer, in favour of retaining his identity as a reflective agent, which he cannot shed.

Here it is worth distinguishing between reflection and rationality. What is the difference between being "re-

flectively unstable" and being irrational? Earlier I distinguished between irrationality and inconsistency; to be irrational is to act in a way that is contrary to what one has reason to do, and to be inconsistent is to act on a particular reason one has which is in conflict with another reason one has for acting in another way. Caligula is not irrational in this sense, because he has reason to act in the way that he does, i.e. to torture, because he values the identity of a sadistic torturer, so he is acting in line with what he has reason to do. Caligula is, however, inconsistent, because he is acting on reasons which arise from his practical identity as a sadistic torturer, and these reasons conflict with the reasons that arise from his identity as a reflective agent. This means Caligula is also "reflectively unstable". So we cannot convict Caligula of irrationality, but we can convict him of inconsistency.

Here we see how Korsgaard accommodates Williams' view, as no Korsgaardian reflective agent could really be irrational, since an agent can only act in a particular way if she has some reason to do so, and so no reflective agent can act in a way that is contrary to what she has reason to do. This is closely tied to Williams' internal-reasons view. Williams has argued against the hope that morality is closely connected to rationality. This is because, for Williams (1981), an agent A has reason to φ if and only if A could reach the conclusion to φ by

a sound deliberative route from the motivations that A actually has in her actual motivational set S — i.e. the set of her desires, beliefs, attitudes, etc. Korsgaard builds this into her argument from the start, as the defining characteristic of Korsgaard's moral agent is that she is essentially reflective, that is, she can only act if she has some desire to act in some way (i.e. she can only act if she has something in her motivational set) and if she reflectively endorses one desire over another (i.e. through a sound deliberative route she concludes in favour of φ-ing). Because Caligula is a reflective agent, he has 'valuing humanity' in his motivational set, and so there is a sound deliberative route that he could take to reach the conclusion that he should not torture, so he has an internal reason not to torture. Of course, Caligula also has reason to torture, because he endorses the identity of a sadistic torturer, so he is "reflectively unstable" and inconsistent.

Korsgaard also accommodates McDowell, a reasons-externalist, here, who wants to say that Caligula should be convicted of something more than merely being nasty, for he has failed to see and act on the deeper, moral reasons. Whilst Korsgaard's view does not allow us to convict Caligula of irrationality per se, it does allow us to convict him (and, indeed, any other agent who endorses a practical identity that is in and of itself contradictory to the value of humanity) of being

"reflectively unstable", of being inconsistent. FitzPatrick puts this in more precise terms, writing that, just like one who holds 'P or not-Q and not-P' is also committed to holding 'not-Q', Korsgaard's reflective agent is committed to the value of humanity whenever she exercises any kind of agency at all, that is, whenever she acts at all (FitzPatrick 2005, 672), regardless of whether or not she has taken that sound deliberative route, regardless of whether or not she has reflected, because if she were to reflect at all, she would come to that conclusion (that she is committed to the value of humanity and thus has moral reasons). This position satisfies (some of) the demands of the reasons-externalist because, though moral reasons are internal reasons, they are also universal — all reflective agents, all human beings, have these reasons in their motivational set. This means that all human beings have reason to act morally.

Critics of Korsgaard have argued that her argument fails to properly answer the sceptic because her conclusion is inescapably conditional, if one values one's own reflective agency, then one has moral reasons. But this criticism is quite misguided in two different ways. First, Korsgaard's conditional conclusion is in fact much stronger than this, it is rather, if a reflective agent acts at all, then she has moral reasons. This antecedent is much more difficult to contest or deny. Second, even if the criticism did point to a weaker conditional conclu-

sion, Korsgaard's argument for it still starts from a reasonably uncontroversial claim about the structure of the human mind, that it is "essentially reflective" (Korsgaard 1996, 92). This antecedent is still not one that the skeptic can easily contest or deny, perhaps especially because Korsgaard is not making claims about the ways the mind or the world really are, but rather about the ways that we experience them. The very moment the skeptic asks 'why should I be moral?' she already reveals that she is a reflective agent and that she experiences the world and her agency in the way that Korsgaard describes. Indeed, as Allan Gibbard writes, Korsgaard starts "from the plight of *anyone* who reflects on what to do and why", even the skeptic who asks 'why should I be moral?' (Gibbard 1999, 140, my emphasis). This makes Korsgaard's argument particularly immune to skepticism, despite its conditional conclusion.

Another, more serious objection to Korsgaard's position is that she makes too far a jump from valuing one's own humanity to valuing the humanity of others. Essentially, why should Caligula value the humanity of his victims merely because he values his own humanity, or that of his friends and family? Korsgaard writes that to value anything that is in contradiction to the value of humanity leads to a pragmatic contradiction. This is because our capacity to value anything comes from our identity as human beings, and so valuing humanity

is what it is to be a human being.

As I have argued above, Caligula is an inconsistent agent because he acts on the reasons that arise from his identity as a sadistic torturer, and these reasons are in conflict with the reasons that arise from his identity as a reflective agent. Caligula feels the pull (or at the very least would do so if he reflected more deeply) in both directions. Korsgaard's point is that in any such situation where an agent has reasons to act in opposite ways, the agent becomes "reflectively unstable". This does not strike me as a particularly controversial claim. Beyond this, when an agent is in such a "reflectively unstable" position, and acts on a reason that is in conflict with reasons that arise from her identity as a reflective agent (as a moral agent), the agent commits herself to both valuing humanity and not valuing humanity.[6] This is a "reflectively unstable" pragmatic contradiction because any agent who is committed to valuing 'P' and 'not-P' commits oneself to a pragmatic contradiction, and it is impossible to act on both of these values, one must choose, and so one is forced to take a sound deliberative route to the conclusion that 'P or[7] not-P', to φ or to not-φ, and if one reflects enough one will shed any identity that is in conflict with the value of humanity.

FitzPatrick frames this issue in terms of 'valuing oneself' or 'seeing oneself as unconditionally valuable', since

one is the source of value[8] (FitzPatrick 2005, 666). So, because we value things, and because we are the source of value, we must regard ourselves as "unconditionally valuable", or "value-conferring" (FitzPatrick 2005, 662-3). He argues that this is the step in Korsgaard's argument which is unjustified, because, according to him, it requires a "psychological necessity" that is simply not satisfied in reality (FitzPatrick 2005, 666-7). I have just shown why it is a pragmatic contradiction to have any value whatever that contradicts the value of humanity, but FitzPatrick takes issue with an earlier step in Korsgaard's argument, the move from our valuing anything at all to our valuing ourselves (i.e. our own humanity). Is it necessary to establish Korsgaard's argument that an agent who has values also believe that she is herself valuable in her capacity to give objects value? In Korsgaard's terms, is it necessary for us to value even our own humanity?

This is a crucial step in Korsgaard's argument to establish that we have moral reasons, but I don't think that this step demands a kind of "psychological necessity" as FitzPatrick understands it. Here, FitzPatrick seems to mean that the agent must believe herself to be unconditionally valuable when she makes any choice whatever. But this is not really the case. Most of the time we are driven by parts of our identity that are not the moral part, that is, most of the time we don't act on moral

reasons, but on reasons that arise from other practical identities we endorse. Just as a student, for example, has reason to get up early and get to class on time. Of course these identities, and therefore the reasons that arise from them, are in a sense secondary to our moral identity, and they must not conflict with our moral identity (otherwise we will find ourselves in a "reflectively unstable" position), and so the value of humanity is implicit in all of our other values, but we need not be, and indeed we aren't, constantly conscious of this fact about ourselves.

I have reason, for example, to get to class on time, and this reason arises from a practical conception of myself as a student, an identity which I value and endorse. I may come to question why it is that I value this practical identity, and I may come to give it up upon reflection. Upon further reflection I might ask myself why it is that I value anything at all, why it is that I have the capacity to value things, and then I will be confronted with certain facts about the structure of my mind, and my most fundamental identity as a reflective, moral agent, and then come to value this in myself, in itself, and in others. But this occurs only under reflection, only when I come to question my capacity to value anything whatever, or, indeed, when I am faced with a difficult choice between different parts of my identity, one of which is the moral, reflective part. Note here the point made above that

Korsgaard's conclusion is inescapably conditional; if I were to reflect, then I would come to the conclusion that I must value humanity and therefore have moral reasons.

The point here is the one made by Williams that an agent A has an internal reason to φ if and only if A could reach the conclusion to φ by a sound deliberative route from A's actual motivational set S. By virtue of acting at all A has 'valuing humanity' in her motivational set, and so there is a sound deliberative route she could take to the conscious conclusion to value humanity, but it is not necessary that she be constantly conscious of this fact about her motivational set. From this we can establish that the "psychological necessity" that Fitz-Patrick identifies is not really a necessity for Korsgaard at all; what is necessary for her is that an agent would come to value herself (and consequently value humanity in general and in others) if she came to reflect upon it.

We are now left with the issues that arise from the Dr. Jekyll-like torturer, who does not endorse the identity of a sadistic torturer but who nonetheless succumbs to his desire to torture others, and the Satyr-like, wanton torturer, who neither endorses nor rejects the identity of a sadistic torturer. Let's begin with Dr. Jekyll.

Dr. Jekyll's is an issue of weak will, and it is a more difficult

and interesting case than Caligula's. Dr. Jekyll resents the fact that there is a sadistic side to him and actively rejects the practical identity of a sadistic torturer, yet Dr. Jekyll still succumbs to his desires to torture others when those desires arise, though he goes on to regret having acted on those desires later. Can we convict Dr. Jekyll of irrationality? It is certainly true that Dr. Jekyll has internal reasons not to act on his sadistic desires when they do arise, but does he also have external reasons not to do so? That is, does Dr. Jekyll, or any reflective agent, have an external reason to always treat humanity as an end? I have argued above that Korsgaard is a reasons-internalist, and so she holds the view that an agent A has reason to φ if and only if A could reach the conclusion to φ by a sound deliberative route from A's actual motivational set S. But, of course, all reflective agents have moral reasons — this is as close as a reasons-internalist can get to a universal, external reason. There are two ways that we can view and resolve the problem from this perspective.

First, we could say that whenever a sadistic desire arises in Dr. Jekyll his motivational set changes so significantly that there is no sound deliberative route he could take to arrive at the conclusion not to act on that sadistic desire,[9] so that he ceases to be a reflective agent at all, and is more like an animal acting on instinct.[10] I don't think that Korsgaard would accept this as a possibili-

ty, for she thinks that the very fact of being a member of the human species means that one is a reflective agent, but even if it were the case that Dr. Jekyll essentially ceased to be a human being (defined relevantly as a reflective agent, and not in the mere biological sense) when he acted on his sadistic desires, we still could not convict him of irrationality. Just as we would not convict an animal of irrationality when it harms or kills another animal, we cannot convict Dr. Jekyll of irrationality in this case, for he does not meet the requirements of reflective agency at all, that is, he is arational, he requires no reason for acting.[11]

The second approach we can take here is to argue that, when a sadistic desire arises in Dr. Jekyll his motivational set changes so as to contain an endorsement of the value of sadism. In this case, Dr. Jekyll essentially becomes Caligula when he acts on his sadistic desires, but there is still a sound deliberative route he could take from his actual motivational set to arrive at the conclusion not to act on his sadistic desires. In this case, we still cannot convict Dr. Jekyll of irrationality, for, as I have shown above, we cannot convict Caligula of irrationality. Although we can (and, indeed, should) convict Dr. Jekyll of inconsistency in this case, just as we have convicted Caligula of the same offence above.

Thus, from the perspective of Korsgaard's neo-Kantian

approach, Dr. Jekyll is either a genuine wanton, that is, an arational creature who does not require reasons for acting, or Dr. Jekyll is more like Caligula, an inconsistent agent. Either way, we cannot convict Dr. Jekyll of irrationality, though we may convict him of arationality or of inconsistency.

We are now left with the issue of the Satyr, the (Korsgaardian) wanton who neither rejects nor endorses the practical identity of a sadistic torturer, but merely takes each and every present desire to be a reason for acting, and occasionally has sadistic desires which he takes to be reasons for action. Note that this is not the same as the first approach to the Dr. Jekyll case above — the Satyr is still a reflective agent in this case, he still requires reasons for action (see Korsgaard 1996, 99). The issue here is one of the domain over which the law that we act on must range.

As I have said above, Korsgaard's approach requires an agent to endorse a particular desire in order to act on it, and in order to do that the agent requires reasons to endorse one desire over another, and in order to do that the agent must act according to a law. This is because the free will of the agent[12] is a "rational causality", meaning it acts according to some law. Because the will is free, it "must be entirely self-determining", but because the will is also a rational causality, it must act

according to some law, for "it cannot be conceived as acting and choosing for no reason" (Korsgaard 1996, 97-8). So the free will "must have its own law", but again we are faced with the same problem as before; how can the will have reason for acting on one law rather than another? Korsgaard concludes that Kant's categorical imperative is the answer to this question, because all it does is "merely tell us to choose a law", "its only constraint on our choice is that it has the form of a law" (Korsgaard, 1996, 98). So, the categorical imperative "describes what a free will must do in order to be what it is", in order to be a free, rational causality (Korsgaard 1996, 98).

But this law is not necessarily a moral one — here the issue of the domain over which the law must range must be considered. The wanton Satyr acts on a law that commands taking each present desire as a reason for acting — this is not a moral law. The moral law must range over every rational being to give rise to moral reasons (Korsgaard 1996, 99). But does the Satyr have reason to act according to the moral law rather than the wanton law? Is there a sound deliberative route from his actual motivational set that he could take to arrive at the moral law? Korsgaard argues that there is.

Korsgaard writes that "the reflective structure of the mind is a source of 'self-consciousness' because it

forces us to have a conception of ourselves" (Korsgaard 1996, 100). This means that when we choose to act on one desire over another we experience "something over and above" our desires which chooses between them, and that is what we experience as 'ourselves' (Korsgaard 1996, 100). This, in turn, means that the law which we choose to act on must be one that we regard as expressive of ourselves, that is, it must be expressive of our particular practical identities, the most central of which is the practical identity of a reflective agent. So the law that we act on must be representative of this basic fact about who we are, and so the wanton law is not a proper law for a human being. Essentially, according to Korsgaard, no human being could be satisfied with such a law, for it does not properly represent oneself, and it does not properly solve the practical problem which Korsgaard identifies, that is, the problem of how to decide what to do and why — the wanton law fails to answer the 'why' part of the problem in an adequate way.

So not only is there a sound deliberative route the Satyr could take from his current motivational set to conclude that he should act on the moral law, but it is inevitable that he will take such a route if he truly is a reflective agent. Again, either the Satyr continues to behave like a wanton — so that we may conclude that he is a genuine wanton in Frankfurt's sense of the word,

and therefore not a properly reflective agent at all but an arational being —, or the Satyr will, upon reflection, accept the moral law and cease to act on his sadistic desires, and so there will be no need to convict him of irrationality or even of inconsistency. Ultimately, the Satyr is neither irrational, nor inconsistent like Caligula and Dr. Jekyll above. The Satyr is either a genuine wanton, that is, arational, or he will, out of his own accord, begin to act on the moral law.

So, can a sadistic torturer be convicted of irrationality? In this essay I have shown how Korsgaard's approach can give us something of a middle-ground between the reasons-internalism and the reasons-externalism views, allowing us to convict a sadistic torturer of inconsistency or of arationality, though not of irrationality. But what does this mean for the "contemporary form of the Kantian hope" (Mercer n.d., 5) that morality is intrinsically tied to our capacity to reason? I have shown that any sadistic torturer — indeed any reflective agent — has an internal reason to act morally towards others because 'valuing humanity' is part of the actual motivational set of any and all reflective agents. So any and all reflective agents have internal reasons to act morally, though perhaps not overriding reasons to do so. A reflective agent who acts in a way that is contrary to the value of humanity, a "reflectively unstable" agent, is inconsistent. And this "reflectively unstable" position is

likely to encourage an inconsistent agent to reflect, and it is through this reflection that they will come to see that they necessarily value humanity and, therefore, should act morally. So, ultimately, a sadistic torturer cannot be convicted of irrationality. Indeed, irrationality is not a real possibility for a Korsgaardian reflective agent; either one is an inconsistent reflective agent, like Caligula, or one is arational, a genuine wanton, who does not require reasons for acting and is therefore, not a reflective agent at all.

Notes

1. Though McDowell does concede to Williams that 'not seeing the proper reasons' is not necessarily the same as 'irrationality'.
2. Here I use the term 'wanton' in the way in which Korsgaard uses it, rather than the way in which Harry Frankfurt uses it. Korsgaard makes this explicit distinction in *The Sources of Normativity* (1996, 99). Frankfurt's wanton is a genuine wanton, who merely acts on each and every present desire as they arise. Korsgaard's wanton is not a genuine wanton in this way, but a reflective agent who misjudges the domain of the law she acts on. I will go on to discuss this below in relation to Dr. Jekyll and the Satyr.
3. From here onwards I will use the terms 'reflective agent' and 'human being', and the terms 'reflective agency' and 'humanity' interchangeably unless otherwise explicitly stated.
4. Here, Korsgaard is not making a metaphysical claim about autonomy or free will, she is merely describing what she takes to be our experience of ourselves. Whenever we make a choice we feel that this choice was made by us, that we are self-legislating beings and not merely a battlefield of competing desires — this is how we experience choice.

5. This is the issue that arises in the case of Caligula, the agent who values his identity as a sadistic torturer. I will explore this in more detail below.
6. She commits herself to valuing humanity because she is constantly committed to this, by virtue of valuing anything at all. She commits herself to not valuing humanity because she acts on a reason that is in conflict with her moral reasons.
7. This is the exclusive 'or', of course.
8. That is, realism is false in the sense that there are no mind-independent moral properties 'out there'.
9. This is indeed the case in Stevenson's original novella The *Strange Case of Dr. Jekyll and Mr. Hyde*, since Dr. Jekyll literally becomes another person, Mr. Hyde, when his sadistic desires arise.
10. Indeed, human beings often act in this way. For example, when we touch a hot plate that burns our hands we instinctively and unreflectively let go immediately to protect ourselves — it is in this way that Dr. Jekyll seems to act in these circumstances. Of course, in the case of the hot plate our actions do not conflict with the value of humanity and so no problem arises as it does for Dr. Jekyll here.
11. Here, Dr. Jekyll is a genuine wanton in Harry Frankfurt's sense.
12. Again, this is not a metaphysical claim, but merely a claim about how we experience choice.

References

FitzPatrick, William J. 2005. "The Practical Turn in Ethical Theory: Korsgaard's Constructivism, Realism, and the Nature of Normativity." In *Ethics* 115 (4), edited by Gerald Dworkin, 651-91. Chicago: University of Chicago Press.

Gibbard, Allan. 1999. "Morality as Consistency in Living: Korsgaard's Kantian Lectures." In *Ethics* 110 (1), edited by Gerald Dworkin, 651-91. Chicago: University of Chicago Press.

Korsgaard, Christine. 1996. *The Sources of Normativity.* Cambridge: Cambridge University Press.

McDowell, John. 1995. "Might there be external reasons?" In *World, Mind and Ethics*, edited by J.E.J. Altham & Ross Harrison, 68-85. Cambridge: Cambridge University Press.

Mercer, Mark. n.d. John McDowell on External Reasons. Accessed on 10/20/2018. http://professormarkmercer.ca/papers/in_proress/John_McDowell_on_External_Reasons.pdf.

Williams, Bernard. 1981. "Internal and External Reasons." In *Moral Luck*, 101-14. Cambridge: Cambridge University Press.ress/John_McDowell_on_External_Reasons.pdf.

Ontic Vagueness in Temporary Existence: A Challenge to Sullivan's Minimal A-Theoretic Metaphysics of Time

Jordan Elizabeth Bridges

Introduction

There are two major theories of time that aim to give an answer to what time is like: A-theory and B-theory. Roughly, for the A-theorist, time is real, time passes, and there is an objective, metaphysically privileged present. (Whereas for the B-theorist, time is merely another dimension, time does not pass, and there is no objective, metaphysically privileged time.) A-theorists assign properties like, "X is past," or "X is present," while B-theorists ascribe relations like, "X is before Y, or "X is simultaneous to Y." A-theoretic properties will change, because which time is objectively present will change. B-theoretic relations will not change, because time is just another dimension with no objective, metaphysically privileged present.

In the paper "The Minimal A-Theory," Meghan Sullivan outlines a version of the A-theoretic model of time that does not include temporaryism, the view that there are

temporary existents. Sullivan argues that temporaryists have not succeeded in capturing a commonsensical belief about existential change. Absent this kind of intuitive backing, Sullivan asserts that we should accept the more metaphysically advantageous alternative to temporaryism: permenatism, the view that everything always exists.

Sullivan's minimal A-theory is unorthodox, but not immediately implausible. Most A-theorists are also temporaryists, but A-theory is perfectly compatible with permenatism. To show that temporaryists have not been successful in their endeavor to best capture commonsensical belief, Sullivan points to a vagueness in our ordinary beliefs about change. In this paper, I will survey Sullivan's starting assumptions and her argument for determinate temporary existence, making explicit her criteria for possible vagueness in our beliefs. I will then raise concerns with a premise in her argument against the Moorean argument for temporaryism. I will do so by articulating how one model of ontic vagueness challenges Sullivan's argument that temporaryism does not have a Moorean advantage. Ontic vagueness, or metaphysical indeterminacy, is simply a kind of vagueness in *what there actually is* rather than in our *descriptions or knowledge of what there is*. Ontic vagueness meets Sullivan's criteria for the kind of vagueness that could be present in the temporary ex-

istential sentence, which is just the logic-ese used to capture the claim that some objects change with respect to existence. Finally, I will consider some of the ramifications of a successful ontic vagueness challenge, suggesting how best a non-temporaryist A-theorist might resist this challenge.[1]

Sullivan's Minimal A-Theory

Sullivan begins by assuming A-theory. For Sullivan, this assumption entails assuming:

> FUNDAMENTAL TENSE: There is a fundamental distinction between the present and other times, and expressing this distinction requires primitive tense operators like "it was the case that..." (usually abbreviated with P), "it will be the case that..." (F) or "it is always the case that..."(which I will abbreviate with □).

> A-PROPERTY CHANGE: Objects do not require temporal parts or time-relational properties to undergo change. Some objects have temporary non-relational properties and endure through change. Using the "always" tense operator, we can express the view most perspicuously: For some property C, ∃x(C(x)& ¬□C(x)).[2]

To account for how objects persist in or change through time, some philosophers hold that objects have temporal parts. Temporal parts are the subject of the incompatible properties involved in change, and the composite of these temporal parts is the object which persists over time. Temporal parts are a somewhat imperfect analog to spatial parts. My younger sister's beloved ombre slippers have spatial parts because they vary in color across space. Likewise, persisting objects vary across spacetime. For instance, the cold pot of water from ten minutes prior to my typing is now a boiling pot of water, but the pot of water is not itself subject to the incompatible properties being cold and being boiling; rather, the earlier temporal stage of the four-dimensional object called the pot of water is the subject of the property being cold, and the current temporal stage of the object is the subject of the property being boiling. Other philosophers think that temporary properties, like being cold or being boiling, are really relations to a time, where a time-relational property would be something like, "In relation to the present time, the pot of water has the property being-boiling."

Sullivan also assumes a neo-Quineanism she characterizes as follows:

> UNIVOCAL EXISTENCE: There is a single, fundamental sense of "exists" of interest to metaphysics, and it is denoted by the existential quantifier.[3]

This assumption states that there is an answer to the question of what change is fundamentally. For

neo-Quineans, "The debate about change is substantive if we can translate different theories of change into logic-ese and show that they must quantify over different domains."[4] Finally, she articulates the view she will aim to refute and the view she proposes neo-Quinean A-theorists accept in its place, respectively:

> TEMPORARY EXISTENCE: Some objects change with respect to existence. In logic-ese we express this using what I will call a bare existential sentence: ∃x¬□∃y(x=y). The sentence is bare because the only predicate it uses is absolute identity.
>
> PERMANENT EXISTENCE: Everything always exists: ∀x□∃y(x=y).[5]

Univocal existence means that the A-theorist cannot describe change in existence as merely a property change. Because the A-theorist needs more tools to describe change in existence, she might opt to accept temporary existence. Temporary existence is commonly thought to have an advantage over permanent existence because there is a good Moorean argument for it. This means that most A-theorists, according to Sullivan, believe that temporaryism is so clearly supported by common sense that absent quite strong reasons for the contrary, it would be irrational to deny the view. To challenge the Moorean argument for temporaryism, Sullivan challenges its first premise:

> ENTAILMENT PREMISE: Highly plausible, common sense beliefs entail some P. More specifically: There is a set of natural language sentences ME that express highly plausible beliefs about a certain domain, there is a set of sentences ML that are appropriate logical paraphrases of ME, and ML entails P.[6]

Her argument goes roughly as follows: neo-Quinean A-theorists are committed to the belief that change in existence is "always, necessarily a determinate matter;" however, the ordinary way we talk about creation and destruction involves penumbral states; so our ordinary beliefs fail to track temporary existence, and so the Moorean argument for temporaryism falls apart.

To demonstrate that common sense belief fails to support determinate existence, Sullivan invites us to consider our intuitions on creation and destruction. I'll offer my own example: if I place a wax statue of Elvis Presley into a furnace, I will observe Elvis melting, gradually transforming into a pool of shapeless wax. Before the melting process, the statue of Elvis Presley existed, and after melting, the statue no longer exists. If one were to ask someone when Elvis had ceased to exist, it seems plausible that she could say that there is not an exact time or stage of melting in which Elvis leaves the building. Already, the language I used to describe this slow melting process hints at a vagueness. As he melts in the

furnace, Elvis is in the process of destruction, a penumbral state where, plausibly, it is not determinate whether or not Elvis exists. This example is meant to show that common sense does not always entail determinate existence, because in Elvis' case, common sense might lead us to believe that going out of existence is a gradual, vague process. If destruction is characterized by temporary existence, then common sense about Elvis doesn't always entail determinate existence. In fact, it looks like common sense sometimes entails nondeterminate temporary existence.

For Sullivan, temporary existence is captured by the bare existential sentence, $\exists x \neg \Box \exists y(x=y)$, so if temporary existence is going to be susceptible to vagueness, the vagueness would lie in this sentence. Sullivan considers two prominent theories of vagueness: semantic and epistemic vagueness. Supervaluationists think indeterminacy is a symptom of semantic indecision. A sentence is indeterminate if and only if it has a vague term and the sentence is true on one precisification and false on another. For example, the claim, "This pile is a heap," could be vague because the word "heap" has multiple candidates for denotation. A heap of sand could be anything greater than exactly 1000 grains, but it also seems plausible that a heap of sand is anything greater than 5000 grains. For the epistemicist, vagueness stems from arbitrary extensions fixed by our language. On this view, there is an answer to whether or not something exists, and this answer is fixed by the way we fix the extension of a term, it just might be difficult or impossible for us to discover the precise boundaries of a vague term. For the epistemicist, there is an answer to when grains of sand become a heap, we just can't discover the extension of "heap of sand."

So vagueness on these views is largely a matter of having too many good options for the denotation of a term. Both supervaluationists and epistemicists accept:

> MULTIPLE CANDIDATE DENOTATIONS: A sentence is indeterminate only if there are multiple candidate precise denotations for at least one of its terms and we cannot know which, if any, particular denotation is fixed by linguistic practice.[7]

However, Sullivan finds it implausible for $\exists x \neg \Box \exists y(x=y)$ to have too many candidates for denotation, because each term of $\exists x \neg \Box \exists y(x=y)$ has exactly one denotation. Recall that neo-Quineans think existential quantifiers pick out the single, fundamental sense of "exists," and for this reason, they should hold that "there is no indeterminacy in the quantifier expressions or their attendant variables," furthermore, "Negation is a logical constant — no room for indeterminacy here. All A-theorists are fundamental tensers, so they think that tense operators like $\Box$ have a single denotation — no room

for indeterminacy here."[8] Finally, Sullivan does not find it plausible for identity to be the source of indeterminacy, because to hold this view one would have to reject the A-theoretic account of change. Basically, the argument is that because we have assumed neo-Quineanism and an A-theory, we should not think that there is anything in the temporal existential sentence that we can identify as the source of vagueness.

However, there is a type of vagueness that does not accept the multiple candidate denotations principle: ontic vagueness. Ontic vagueness, or metaphysical indeterminacy, is simply a kind of vagueness in what there is rather than in our descriptions or knowledge of what there is. Sullivan gives this view a brief treatment, "Sojourners on this less-travelled route to indeterminacy maintain that a semantically determinate and epistemically scrutable sentence can nevertheless pick out a state of affairs such that it is indeterminate whether that state of affairs obtains."[9] Sullivan does not offer an argument against ontic vagueness in ∃x¬□∃y(x=y), instead she says, "Here I have little to offer beyond noting that I don't see how [ontic vagueness could obtain here]...purely fundamental facts either obtain or they do not — that's just part of what it is to be a fundamental fact."[10] In other words, fundamental facts about existence may either be true or false. However, there is a model of ontic vagueness that meets these criteria.

Sullivan acknowledges that readers with sympathies towards certain models of ontic vagueness will likely find her response question-begging, but she thinks those who have accepted her preconditions shouldn't be concerned about the possibility of ontic vagueness obtaining. I'll consider her argument more carefully, then I'll outline a model of ontic vagueness that I think causes her treatment most trouble. Finally, I will explore the ramifications of a successful challenge via ontic vagueness to her refutation of the necessarily determinate existence premise.

The Ontic Vagueness Challenge

Sullivan's argument goes like this: fundamental facts are determinate, neo-Quineans hold that existence is fundamental, A-theorists accept fundamental tense, so "neo-Quinean A-theorists should think facts about bare temporary existence are fundamental. The temporary existence principle makes a kind of bare existential claim."[11] Sullivan acknowledges that there may be some fundamental indeterminacy with regards to really weird temporary existence, like quantum objects, but asserts that we shouldn't be concerned with these sorts of cases in a defense of Moorean advantage because quantum physics isn't exactly the stuff of common sense. Finally, she, perhaps rightly, observes that accepting ontic vagueness is not a solution most A-theorists could swallow.

But let's say some A-theorists aren't unhappy with a good model of ontic vagueness, if such a model exists. Sullivan doesn't take ontic vagueness seriously in part because she seems to think that it's not the sort of view of which one can make sense. In her paper "Ontic Vagueness: A Guide for the Perplexed," Elizabeth Barnes offers a model of ontic vagueness that has become increasingly respected for its ability to make sense of metaphysical indeterminacy. I mention the view's prominence not to motivate the view by an appeal to Barnes' authority, but to suggest that if Sullivan cares about addressing and responding to challenges from the most prominent models of vagueness, she ought to consider a well-regarded model of ontic vagueness. I will briefly sketch what a metaphysical indeterminist adopting Barnes' model might take issue with in regards to Sullivan's account.

The following is a general model of Barnes' ontic vagueness:

> (OV) Sentence S is ontically vague iff: were all representational content precisified, there is an admissible precisification of S such that according to that precisification the sentence would still be non-epistemically indeterminate in a way that is Sorites-susceptible (as in, susceptible to the Sorites paradox.)[12]

(OV) is a counterfactual that holds that if a claim is vague, but it isn't vague in its semantics nor is it epistemically vague, then claim is ontically vague (or metaphysically indeterminate).[13]

Furthermore, Barnes can provide a formal translation of (OV):

> (OV*) ∇op at w iff $\exists x(\nabla Ixw \ \& \ x \Rightarrow p) \ \& \sim \exists y(Iyw \ \& \ y \Rightarrow p)$

In English, (OV*) means, "P admits of ontic indeterminacy when x makes p true, but it's indeterminate whether x exists at w. What it takes to make p true is settled, but it's unsettled whether what it takes to make p true obtains."[14]

I leave Barnes to completely defend whether her definition successfully distinguishes itself from vagueness of the semantic or epistemic sort, and whether it develops adequate constraints for a model of ontic vagueness. I don't seek to provide another argument for this model of ontic vagueness; I am merely offering a sketch of how a proponent of this view could argue that this ontic vagueness model meets the criteria Sullivan articulates for an adequate model of vagueness.

Recall that Sullivan attempts to exhaust "the options for

explaining any indeterminacy in temporary existence by appeal to multiple candidate denotations for a term," and not unsuccessfully, I think.[15] In the background to my response, I outlined Sullivan's reasoning in her rejection of the possibility of $\exists x \neg \Box \exists y(x=y)$ being semantically or epistemically vague. However, as Sullivan notes, not everyone who thinks $\exists x \neg \Box \exists y(x=y)$ is vague will be content to accept multiple candidate denotations. The Moorean assumption Sullivan is challenging is that this sentence is vague. If one accepts Barnes' model, the conditions for vagueness have been met. If the conditions for vagueness in $\exists x \neg \Box \exists y(x=y)$ were met, then Sullivan has not yet succeeded in breaking the stalemate between temporaryists and permenatists.

But first, we must see how Barnes' model meets the conditions for the right sort of vagueness to cause problems for determinate existence. The sentence $\exists x \neg \Box \exists y(x=y)$ is ontically vague because all representational content has been precisified and there is an admissible precisification of S such that according to that precisification the sentence would still be non-epistemically indeterminate in a way that is Sorites-susceptible (see OV). As Sullivan suspected, someone who accepts a model like Barnes' as credible will find Sullivan's dismissal of such models question-begging.

To give an example of how ontic vagueness might work

here, let's pretend that facts about melting snowmen or wax Elvises are the sorts of fundamental facts Sullivan is concerned with; they do not depend on our understanding of what count as snowmen or wax Elvises, but they are deep facts about the laws of the universe. If this seems immediately objectionable, then one should take issue with Sullivan's use of these examples as well. It could be the case that it is metaphysically indeterminate whether Elvis exists. On Barnes' model, it is not the case that Elvis existing and Elvis not existing are equally good candidates for what is going on in the actual world. Rather, determinately, only one of these candidates is the best, it's just indeterminate which is actualized, representing the actual world as an ersatz possible one. Determinately, only one possibility is actualized, and determinately the actualized possibility is either that Elvis exists or Elvis does not exist at time t. This analysis maintains a bivalent model of indeterminacy, unlike the "multiple good candidates" principle proponents of other forms of vagueness might endorse.

One might get the sense from the example I used to illustrate how ontic vagueness works that if Sullivan wanted to defend her dismissal of ontic vagueness, she ought to hold that such vagueness just isn't an appropriate candidate in the cases of determinate existence she cares about. Recall that Sullivan cares about fundamental facts that don't depend on observers, con-

ventions, and the like, and what counts as a legitimate snowman or Elvis statue is likely not going to count as a fundamental fact on her view. But the proponent of ontic vagueness does not need to give examples of ontically vague things. Barnes' ontic vagueness isn't a positive definition as such; rather, it is a counterfactual that obtains if and only if other options for vagueness have been exhausted. For this reason, defenders of this view need not provide examples of a plausible metaphysical indeterminacy within the realm of the commonsensical (unlike that of quantum physics) but still metaphysically significant. Metaphysical indeterminacy obtains when there's vagueness and we can't find a better option to blame the vagueness on. In other words, we if we have good reason to think that there is a vagueness in the temporary existential sentence and we have good reason to accept Barnes' model of ontic vagueness, then we can say that the temporary existential sentence is vague, even though we cannot point to where the vagueness is in the logic-ese. The dialectical background began with the assumption that our common-sense beliefs entailed indeterminacy with regards to existence. For this reason, the common-sense belief should lead us to hold that there is an ontic vagueness in $\exists x \neg \Box \exists y(x=y)$, because common-sense belief entails indeterminacy and ontic vagueness is the remaining option after semantic and epistemic vagueness have been dismissed. Not only is ontic vagueness

merely the last remaining option, but it actually obtains by definition precisely because all other options have been exhausted.

Conclusion

If one wanted to continue to maintain that $\exists x \neg \Box \exists y(x=y)$ is vague in light of Sullivan's critiques of semantic or epistemic vagueness here, accepting Barnes' model of ontic vagueness is her best bet. If one accepts Barnes' view of ontic vagueness (and such a person would be the sort Sullivan would have to address in considering the third prominent form of vagueness) then one should agree that ontic vagueness obtains in $\exists x \neg \Box \exists y(x=y)$. If ontic vagueness obtains, then we should be able to reject the second premise in Sullivan's reductio of the Moorean argument: necessarily and always, temporary existence entails determinate temporary existence. Without this assumption, we cannot get the contradiction which thwarts the Moorean conclusion that there are worlds and times where determinate temporary existence is both true and false.

It now appears that the best way for Sullivan to thwart the ontic vagueness challenge would be to challenge ontic vagueness itself, showing that it is somehow incompatible with the combination of assumptions entailed by A-theory and neo-Quineanism or that the model doesn't hold up for some other reason. I suspect this would be quite the challenge, because

Barnes' model really seems to align with neo-Quinean standards (as I've just attempted to show) and I can't see how fundamental tense nor A-theoretic change could cause problems for the view. All that remains is for the permenatist seeking to challenge the Moorean argument in the manner Sullivan does to criticize ontic vagueness as incompatible with some other commonly-held belief, internally inconsistent, or otherwise troublesome in some way. The purpose of this paper is to show how one ontic vagueness challenge could push on Sullivan's argument, allowing temporaryists to defend their Moorean advantage. For this reason, I will leave defending Barnes' model to others. Because ontic vagueness is a credible option for maintaining that $\exists x \neg \Box \exists y(x=y)$ is vague, the Moorean stalemate remains as it was at the start of the Sullivan paper. However, some interesting progress has been made. The discussion shows that given certain presumptions, ontic vagueness is sometimes entailed by our commonsense beliefs. Such a result is not insignificant, considering how strange and unintuitive ontic vagueness may at first seem.

Notes

1. Thanks to Elizabeth Barnes and Ross Cameron for comments on earlier drafts of this paper.
2. Sullivan, Meghan. "The Minimal A-Theory." *Philosophical Studies* 158, no. 2 (2012): 149–74. https://doi.org/10.1007/s11098-012-9888-5. 150-151.
3. Ibid. 150.
4. Ibid.
5. Ibid. 152.
6. Ibid. 154.
7. Ibid. 161.
8. Ibid.
9. Ibid. 161-162.
10. Ibid. 162.
11. Ibid.
12. Barnes, Elizabeth. "Ontic Vagueness: A Guide for the Perplexed." Noûs 44, no. 4 (2010): 604.
13. Barnes' model leaves open the possibility that credible models of some other form of indeterminacy could be developed; her counterfactual could be modified to account for these new options. For justification of the adequacy of the use of a counterfactual, see Barnes (2010).
14. Ibid. 609.
15. Sullivan, Meghan. "The Minimal A-Theory." *Philosophical Studies* 158, no. 2 (2012): 149–74. https://doi.org/10.1007/s11098-012-9888-5. 161.

References

Barnes, Elizabeth. 2010. "Ontic Vagueness: A Guide for the Perplexed." *Noûs* 44, no. 4: 604.

Sullivan, Meghan. 2012. "The Minimal A-Theory." *Philosophical Studies* 158, no. 2: 149–74. https://doi.org/10.1007/s11098-012-

9888-5. 150-151.

Listening to Socrates: Reevaluating Stephanus 327, Establishing Prefigurative Analysis, and Performing Dianoesis in Plato's Republic

Bradley Davis

Book I of the *Republic* presents a number of problems and a wealth of information for Plato scholars. Historians can provide general readers with a sense of what Plato's contemporaries might have understood from it, as the characters introduced have real historical referents whose backgrounds seem suggestive for the *Republic*. Classicists have noted questions of transmission for Book I; stylometrically, it does not fit with later parts of the work and may have started as a separate dialogue from the larger *Republic* or a proto-*Republic*. Yet, the literary nature of Book I and its consequences for interpreting the remainder of the work have been insufficiently explored. I intend to focus my study on the style of Book I insofar as it influences interpretation of all that follows in the Republic with an emphasis on prefiguration in Stephanus 327. Dramatic prefiguration, as described in the *Republic* by George Rudebusch, "is the literary device, found in Greek tragedy, of using an

image at the beginning to represent or prefigure ideas developed later in the work" (Rudebusch 2002, 77). While claims of prefiguration have been made previously, Plato scholars generally seem reticent to accept the notion. This is understandable considering that many prefigured interpretations do not well adhere to the text of the *Republic*. Rather than use prefiguration in a balanced hermeneutical process, well-known scholars seem to read their preferred interpretations into Book I — a mistake I will try to highlight and correct. In this paper, I will demonstrate the importance and utility of prefiguration in Republic Book I, show how it may be employed for *interpretative* insight, and suggest a path forward for prefiguration in Republic scholarship.

I will briefly discuss what prefiguration is and how it is valuable for *Republic* interpretation. To demonstrate what is at stake and the extent to which *Republic* scholarship is dependent on Book I, I will provide a heterodox reading of 327 that seeks to reframe the *Republic* around a question of strength. Regardless of this reading's merits, I hope it will provoke reconsideration of Book I details and their consequences for greater interpretative claims. Subsequently, I will discuss general problems with prefigurative interpretation and show where previous scholarship makes unsubstantiated or weak prefigurative claims. I am not certain that there are any criteria sufficient for adjudicating the veracity

of prefigurative arguments but invite scholars to push against Book I readings, new and tired, with a view towards how they shape the remainder of the *Republic*. As we will see, establishing dramatic prefiguration is crucial for any strong, comprehensive study of the Republic. Even further, establishing dramatic prefiguration is to perform dianoesis, Socrates' method of reconciling faulty images in order to progress further towards knowledge of the forms — perhaps, to knowledge of any sort. This epistemological method is not only the key to a philosopher's education but to any reader's hermeneutic for the *Republic*.

I

Dramatic prefiguration holds that elements of Book I are used by Plato to foreshadow or indicate thoughts that will be developed later on in the Republic. As such, it is only useful if one determines that Book I was intended by Plato to be included with the *Republic* as it exists. Rudebusch notes that dramatic prefiguration was a common literary device in Greek tragedy, and provides the example of Aeschylus' Agamemnon:

> At the beginning of the play we are told of two eagles, one black, one white- tailed, who, in full view of the army, devour a pregnant hare with all her unborn young (lines 111– 120).

This image prefigures the main action of the play, the

murder of Iphigenia by two kings, her father and uncle, who sacrifice her, "stopping her from her course" before the birth of children (Rudebusch 2002, 77). The form of Plato's dialogues is important for the philosophic content. In Book VII, Socrates tells Glaucon that he cannot simplify the path to knowledge; it requires a journey of reconciling images towards knowledge — similar to the dianoetic method of the Divided Line. Recognizing the flaws or strengths of images is what enables a thinker to move towards recognition of knowledge itself. Socrates' method is to demonstrate that his interlocutors premises are faulty or incomplete, enabling subsequent refinement. Such a technique is employed throughout Book I as Socrates complicates each interlocutor's concept of justice. Through steady development, justice is better defined until Socrates begins the discussion of justice that occupies the remainder of the work, itself growing more complex and nuanced. Socrates' teaching could not occur via doctrine or treatise but only through a series of images.

This style of Socratic argumentation is present throughout the *Republic* and, combined with the dramatic style of the dialogue, makes the adoption of dramatic prefiguration intuitive. Socrates challenges interlocutors with an allegory or image of a concept he is trying to explain that seems to be defective but is heuristically useful. The tripartition of the soul hardly seems to be an ex-

act description of human motivation and behavior, but it is suggestive for a human psychology. The mapping of this tripartite soul to different types of political class is not exact either but provides useful understandings of how different societies might function. Likewise, the segments of ascent within the Cave or Divided Line are not equivalent metaphors, but the two in conjunction provide greater understanding for Plato's theory of knowledge. It would make sense that Plato writes to his readers as Socrates speaks to his interlocutors, with different sections of the work resembling one another while not being exactly equivalent. This inequivalence should be stressed, as I do not believe that the contents of Book I should be understood as perfect images of arguments made later in the work — rather, they are imperfect imitations like all images.

When re-reading the *Republic*, most readers are likely struck by the opening of Book I when Socrates and Glaucon are at the festival of Bendis. Socrates speaks of traveling down to the Piraeus, Polemarchus mentions performances with equestrians holding torches, and Socrates says that he and Glaucon are seeking to return to Athens — to ascend from the harbor. All of these images seem to be symbolic, representing a philosopher guiding a pupil out of the Cave. These dramatic images provide the easiest examples of prefiguration, although I will later show that interpreting what

they prefigure is controversial. In the sections that immediately follow, I will try to provide a close reading of the opening to the *Republic* in order to suggest that overlooked details may have significant interpretative consequences for the work.

II

In his portrait of philosophy and polis, Plato's art is nowhere more evident than in his beautiful opening: Setting up a dramatic stage for his teacher, who is recounting a tale to an unknown audience. With prefiguration in Stephanus page 327, careful readers can glean so much more than simple exposition for the *Republic*. 327 is where Socrates captures his audience and implores them — us — to listen.

I will examine three elements of 327: the ascent, justice of the strongest, and journey from both the beginning of the text and the home of Cephalus onwards. My intent is to show how dramatic elements provide a basis for reimagining and understanding what follows in the *Republic* via prefiguration. The style of Book I is unique in a way that should pique readers' attention. While the whole of the *Republic* does contain dramatic characters who speak with one another and perform some actions, Book I is the only section that seems to have a true dramatic structure; it has significant action and unfolds somewhat like a play. These elements of Book

I have not been sufficiently explored — especially beyond niche interests in political philosophy. I will also suggest corrections to some of the prefigurative literature that does exist. Regardless of these suggestions, I hope to demonstrate that future discussions of the *Republic* would benefit by inspiration from and reconciliation with Book I.

The central thesis of the *Republic* is often debated: should the treatise be considered a work of political science, moral psychology, or philosophic education? Readers with a good memory may recall that while Socrates has much to say on each of these topics, the initial and perhaps central challenge is to Socrates' autonomy: Polemarchus orders Socrates to halt and cease his return home (*Republic* 327b). Socrates went down to the harbor at Piraeus with Glaucon, son of Ariston and brother of Plato, to pray and observe. The goddess Bendis was to be celebrated for the first time and, having seen the Piraens and Thracians perform their rituals, the two journeymen set off to return to Athens before their interdiction. Thus, Polemarchus lives up to his namesake — first-for-fighting, the initiator of conflict in the *Republic* (Rudebusch 2002, 78).

Socrates holds no desire to remain, responding to his momentary captor:

> Polemarchus said, "Socrates, I guess you two are hurrying to get away to town." "That's not a bad guess," I said. (327c)

Socrates seems irritated by Polemarchus' arrest and requests emancipation. Polemarchus demurs and not-so-subtly threatens Socrates. Polemarchus subsequently makes suggestion of Socrates' ignorance of the festivities to follow, and Socrates flippantly responds about the novelty of the festivities. Polemarchus commands Socrates to remain throughout the night and Glaucon again acquiesces. At no point in the exchange does Socrates make a decision of action, accepting or rejecting Polemarchus' decrees. Glaucon is the one who always responds: "Of course we'll wait," "There's no way," "It seems we must stay" (327b-328a). Socrates resists and never assents to his captor, although he does comply. He expresses a desire to be on his way back to Athens. Before any determination has been made as to what justice is, a Socratic conception of justice has been violated. If the ascent from Piraeus in any way represents the acquisition of knowledge, then Socrates' existential desire has been violated — he has been pulled from philosophy and curiosity to the home of Cephalus, from elysium to the polis.

Why might Socrates not want to be drawn into Cephalus' home and into the ensuing debate? Socrates best offers evidence for this and his cagey behavior in Book

VII, when Glaucon affirms Socrates' explanation that:

> "those who have been allowed to spend their time in education continuously to the end... they won't be willing to act, believing they have immigrated to a colony on the Isles of the Blessed while they are still alive?" (519c)

Socrates certainly does not seem to desire to act or engage with the Book I interlocutors. Perhaps, even the agora-minded philosopher becomes caught up in the joy and excitement of his thoughts. There is, after all, a reason why Aristophanes lampoons Socrates. He is fond of intellectual engagement and sees little cause for other activity. One can scarcely imagine Socrates governing Athens. Still, it is curious that Socrates would avoid philosophizing with Cephalus and his sons, unless their time would be spent otherwise. Though, Socrates does mention that:

> "Then it's impossible," I said, "that a multitude be philosophic." ... "And so, those who do philosophize are necessarily blamed by them."(494a)

Is it possible for Cephalus, Polemarchus, and Thrasymachus to philosophize with Socrates? Do they even desire to do so? Perhaps not. Cephalus retreats as soon as he is bested by Socrates, who dispels Cephalus' concept of justice, and proceeds to pray despite

the exchange. Polemarchus appears more interested in Simonides than philosophy proper. Thrasymachus fades out of discussion after growing frustrated with his perception of Socrates manipulating the weaker argument for the stronger. Here, too, Glaucon endorses and bears forward what others would rather dismiss by forcing the conversation to continue. If Socrates is uninterested, it must be out of an unwillingness to return to Greece from the Blessed Isles of his thought and to be accosted by Polemarchus and company.

Why do Polemarchus and the other interlocutors desire discourse with Socrates? They must know that it is to their benefit that Socrates, perhaps Athens' strongest intellect, guides them. While they desire for Socrates to found or lead conversation, Thrasymachus accuses him of trickery despite knowing well that this is Socrates' modus operandi:

> "I certainly believe it," [Thrasymachus] said, "so that Socrates can get away with his usual trick; he'll not answer himself, and when someone else has answered he gets hold of the argument and refutes it." (337e)

Socrates refutes Thrasymachus into submission before ever asserting his own concept of justice, but Glaucon insists on the need to continue. The argument from Thrasymachus seems to be dismissed. Socrates

has persuaded each man that his concept of justice is faulty.

It is odd that Socrates' arguments are generally accepted by his interlocutors. That is, even if they were not convinced that Socrates was correct; they seem to believe that they were proven wrong whilst maintaining their previous behavior. What has Socrates done? Polemarchus confesses that Socrates should not expect being listened to. Polemarchus adequately predicts their unwillingness to listen in confrontation: Cephalus fleeing, Polemarchus being rebuffed and overtaken by Thrasymachus, Thrasymachus frustrated in the corner — at least, for the remainder of Book I. Even beyond closing their ears to Socrates' rebuttals, they certainly do not hear what Socrates seeks: not a manifestation of justice but the thing-in-itself. So, what then does Socrates achieve?

Polemarchus' initial challenge to Socrates is an appeal to force and number,

> "...Do you see how many of us there are?"
> "Of course."
> "Well, then," [Polemarchus] said, "Either prove stronger than these men or stay here." (327b)

Polemarchus challenges Socrates to prove stronger

than the mob. What Polymarchus' notion of "stronger" entails is unclear, and the extent to which the term is ambiguous later in the work has been discussed widely. How Polemarchus could expect Socrates to prove himself stronger is even more obscured, as Polemarchus presumably does not mean through an outnumbered brawl and Socrates later casts a doubt over the possibility of convincing a multitude (493c ff). Some may feel that my reading of the situation is too aggressive. Perhaps, Polemarchus is being more playful or friendly with Socrates than violent. Still, even if the exchange is light-hearted, an insincere threat of violence still constitutes a threat. Further, readers would be remiss to forget that Socrates' died at the hands of a stronger multitude.

Interestingly, this initial exchange precedes Thrasymachus' introduction into the work. Socrates seeks a way out of his arrest aside from competition:

> "Isn't there still one other possibility . . ." I said, "our persuading you that you must let us go?"
> "Could you really persuade," [Polemarchus] said, "if we won't listen." (327c)

Bringing the *Apology* to mind, Polemarchus places Socrates in a bind where he must prove himself stronger in action or in speech — but he cannot and does

not truly persuade a non-audience. What seems to have happened is that Socrates, desiring to leave, embraced Polemarchus' challenge as a means to hasten his departure. The competitive element of the initial debate has been discussed to varying extents, but Socrates does appear to be combating his interlocutors' notions. In Book I his arguments are exclusively negative and defensive, he demonstrates why others' arguments are weak. Socrates does not demonstrate that he or his arguments are strong. Book I ends as follows:

> Before finding out what we were considering at first — what the just is — I let go of that and pursued the consideration of whether it is vice and lack of learning, or wisdom and virtue. And later, when in its turn an argument that injustice is more profitable than justice fell in my way, I could not restrain myself from leaving the other one and going after this one, so that now as a result of the discussion I know nothing. (354)

Here Socrates admits that, rather than seeking to determine what justice is, he sought to break down the others' arguments. No positive argument is established, no strength qua strength is demonstrated as Socrates is on the defense. Rather, the weaknesses of Book I arguments are exploited by Socrates. Though Socrates believes he has earned his release, Glaucon calls Socrates' bluff:

> Now, when I had said this, I thought I was freed from argument. But after all, as it seems, it was only a prelude... [Glaucon said,] "Socrates, do you want to seem to have persuaded us, or truly to persuade us, that it is in every way better to be just than unjust?"

That Socrates considers himself freed from argument, without having to speak or respond in earnest, shows his disinterest in the discussion. Considering that his freedom was seized in order for the arguments to take place, this comment becomes even more striking. From this challenge, the rest of the Republic derives its momentum — yet, the impetus is still Polemarchus' initial arrest of Socrates. Glaucon's comment could suitably be rephrased: Socrates, do you want to pretend to be stronger or prove it? If Socrates had any intention to engage in meaningful philosophical discourse, to learn or to teach, he would not have described his thoughts as above. Socrates provides his account of what justice is, offering a positive and strong account of justice. If he left Piraeus, Socrates proved himself the stronger. Given that Socrates is able to recount his interactions from Cephalus' home, it must be assumed that he succeeded in Polemarchus' challenge.

After all, Plato has Socrates begins the *Republic*'s account of events as follows:

> Socrates: I went down to the Piraeus yesterday... (327a)

When studies of the *Republic* do take Book I into consideration, they often start from Cephalus' household and the characters there within while ignoring this important introduction. While the characters' arguments and historical antecedents are interesting, a more logical and helpful start is the true beginning of the work. While the Bloom translation emphasizes with a colon, the original Greek makes clear that Socrates is the dialogue's reciter. Plato presents him speaking to an unknown audience, most likely himself, Plato, or the book's reader. I find the latter possibility to be most compelling and intuitive but I neither see adjudicating text in the *Republic* nor find much at stake as regards who Socrates' audience is so long as it is acknowledged that the dramatic interlocutors are not who Socrates is actually speaking to. That Socrates is speaking in recitation of the previous day's events has been discussed alongside questions of poetry and imitation within Book X, but not much further. Despite the clear division between narrative and imitative poetry described in the *Republic*, Plato blurs the line when he imitates Socrates providing narration rife with imitative metaphors. Readers may get lulled into the "I said/he said" dialogue of the *Republic*, but Plato's decision of style has consequences for how his thoughts should be considered — particularly regarding poetry.

That Socrates recites the previous day to us, the readers, requires interpretative consequences. Dialogue should proceed between Socrates/Plato and their readers as soon as the long soliloquy that is the *Republic* continues. If this is true, then certain interpretations of the *Republic* must be immediately elevated. One such example is Smith's paper "Plato's Book of Images," which treats Plato's *Republic* as a series of images that readers must reconcile and develop, if readers are participants in a Socratic dialogue then this didactic effort is both reasonable and worthwhile. Scholarship that seeks to integrate Thrasymachus' Argument of the Stronger with the remainder of the work also gains importance with Socrates' demonstration of strength to Polemarchus. Some problems may also stem from this, primarily that Socrates may be an unreliable narrator of the previous day's events from poor memory or otherwise and may be underplaying the arguments of others. This seems particularly true for Thrasymachus, as the negative portrayal of both the character and his arguments could be exaggerated in Socrates' hindsight.

III

It should be clear that Plato took great care in his crafting of the *Republic*, particularly Book I with its dramatic layering and strength as a referent. I believe that the work is best understood as it relates to the initial action of descent/ascent and Polemarchus' arrest of and

challenge to Socrates. Understanding the remainder of the *Republic* as it extends from the dramatic beginning or even from the home of Cephalus will benefit scholarship by grounding textuality and appreciation Plato's literary style — hopefully, interpretations will be able to find prefigured antecedents in Book I. However, some attempts to do this have not been inaccurate due perhaps to confirmation bias in finding antecedents or insufficient reliance on the text, thus discouraging the study of prefiguration.

It is important to note that Socrates' dramatic journey in the *Republic* is his descent to Piraeus and then his arrested attempt and eventual ascent back towards his home in Athens. Brann has emphasized this in her study of the Republic, and she discusses that Books I and X both work as descents into Piraeus as a symbolic Cave and then into Hades, with a super-cave zenith in Books V and VI (Brann 2004). Brann's concentric ring theory and other exposition-heavy interpretations of the Republic often emphasize the first book's descent to the Piraeus. But this is a misplaced focus. Though the descent to Piraeus hangs over the Republic, it is not an action that occurs in the work — akin to the Sphinx's Riddle in *Oedipus Rex*. Book I cannot function as a narrative of descent, as it is about what occurs afterwards. The focus of scholarship should be an inversion of Brann's reading, more akin to Seth Benardete's discus-

sion of the ascent from Piraeus (Benardete 1989). How one reads the direction of this vector should have consequences for understanding not just the characters' journey but also the later images of the Divided Line and the Cave, along with their substantive arguments. Particularly, the notion of what it means to be "like us" as Socrates discusses at 514.

Another image that has been thoroughly discusses it the festival of Bendis, which has been a historical source for studies on Bendis in Athens. Some scholars have noted that the torch race on horseback prefigures the Allegory of the Cave, an example that epitomizes the difficulties of interpreting prefiguration. While it is not clear that Socrates or any of the other characters actually attend the race or any festivities after the initial departure, it is not a parallel presentation of images to the Cave if the horsemen provide both the images from which shadows are drawn and the flame that casts shadows. This is perhaps a quibble, but a much more logical referent for the Cave's images is the part of the festival Socrates did attend. Socrates describes:

> Now, in my opinion, the procession of the native inhabitants was fine; but the one the Thracians conducted was no less fitting a show. (327a)

The description of a show implies entertainment or per-

formance before an audience, roughly but more approximately a Cave-like description. Furthermore, the act of a procession brings to mind a steady march of figures that is much more recognizable as Cave imagery than a horse race. But this also does not purely map onto the Cave allegory with its procession of statues. Leo Strauss has mentioned that as opposed to natural light or the festival's torchlight, the conversations that take place within Cephalus' home are amongst artificial light (Strauss 1997, 64). Rudebusch finds yet another torch in Book I, explaining that the name "Glaucon" should be understood as "gleaming" or torch-like (Rudebusch 2002, 79-80). If taken this way, Glaucon may provide a constant source of shadows no matter how far he and Socrates go in their ascent. It is difficult to determine which one of these Book I images ought to serve as the prefiguration for the procession of images within the Cave, possibly all of Book I is imbued with general reference to these lowliest images and serves as a general prefiguration rather than a particular device. How one determines to best interpret this problem changes what it means to be "like us" and also should change how the scenes of Book I can be reimagined as a Cave.

In *The City and Man*, Strauss also poses Polemarchus' challenge as fundamental to political philosophy. He argues that the *kallipolis* must follow a model of persuasion and compulsion. The philosopher persuades all

others, via the noble lie and otherwise, that the regime is just and ought to be embraced, at least in part as a means of securing the safety of philosophy. Yet, the philosopher does not want to rule the city so the multitude most compel the philosopher to rule as they have been persuaded he or she must. This is exemplified by Book I. Yet, this duality of persuasion and compulsion as a governing principle does not seem to withstand the multitudes' avowed refusal to listen in theory or practice despite Socrates' eventual departure. I admit that Socrates does speak of both persuasion and compulsion in a unified manner, but Strauss's reading does not appear to meet the test of a prefigured Book I. Even with its controversies and my own disagreements with aspects of the work, more attention should be given to *City and Man* insofar as it acknowledges this question of strength in Book I. If the reading of Book I presented above is accepted, prefiguration may develop a different understanding of what Socrates advocates with the *kallipolis*.

As was noted earlier, Socrates is challenged to prove himself stronger than Polemarchus and his companions prior to Thrasymachus' introduction and as a prerequisite to continue the ascent home. I sought to lay a foundation for challenging the bulk of existing literature on Thrasymachus — regardless of whether he and Socrates are engaging in formal debate or agree on a

definition of justice, Socrates seems to adopt Thrasymachus' formulation of justice as nothing but the will of the stronger. Socrates goal amongst the dramatic interlocutors should be best understood as seeking reprieve from his capture, and to develop a theory of justice insofar as he can prove himself the stronger to Glaucon, Thrasymachus, and all of the others. Socrates is successful at this, he is presumably released and Glaucon — alongside the majority of Plato's readers — seem to be persuaded at least in part by Socrates speech. Socrates has subsequently proven himself the strongest philosopher in a long history of thought, not just as the strongest in the Republic. His winning argument, presenting the city-in-speech, provides Socrates, or philosophers like Socrates, the responsibility of governing a city and crafting its laws. Thus, by all of Thrasymachus' metrics, Socrates demonstrates himself the stronger — with justice to his own advantage. But this interpretation seems to devalue the *kallipolis* as an earnest proposal and makes many of Socrates' arguments appear purely instrumental. Separately, if one elevates the importance of Polemarchus' "arrest" of Socrates, then the Republic may draw closer to the *Apology* in providing Socrates with more than just one day to defend himself and the philosophic life. If this interpretation stands and bears fruit, then to read the Republic without the dramatic prefiguration of Book I in mind is to read the Republic erroneously.

~

Greater reference to the prefiguration of Book I in general and 327 in particular will serve to firmly root scholarly discussions in the dramatic text that Plato has provided us. Interpretation rooted in this prefiguration can provide novel and compelling evidence for resolving debates about how to understand later passages in the *Republic*. But prefiguration is insufficient and needs to be one element of a balanced hermeneutic, as it is too easy to read an interpretation into Book I. Nonetheless, the dramatic nature of the *Republic* is too widely ignored, to the detriment of its readers. I believe that insufficient attention has been given to Polemarchus' challenge to Socrates, a reframing of the *Republic* provides alongside this effort by Socrates may not be preferable but should serve to elucidate certain aspects of the work.

Once prefiguration is recognized, readers have a greater ability to comprehend not only the work's discussion of justice but its radical epistemology. As Socrates insists on the importance of dianoesis in discerning knowledge from the images in the world, so too has Plato constructed a work that requires dianoesis to interpret. Book I provides both a microcosm of the work's contradictions and competing truths, while constructing a lens through which the work must be read and reconciled. Plato's *Republic* is therefore both a treatise

on how to understand a concept justice and an exercise in the practice of learning to understand the world of concepts we find ourselves in. The *Republic*'s ability to provide such a deep theory of knowledge within literary elements that in turn interrogate justice has to represent the zenith of philosophy and literature. It demonstrates how inextricable questions of politics are from ethics, ethics from knowledge, and knowledge from the way we reconcile images and descriptions of the world.

Prefigurative scholarship extends beyond just reading Plato's masterpiece, encouraging Socrates' audience to listen. To do otherwise is to ignore the work's beauty and depth, to misunderstand its discussion of justice, and to reject the very method of dianoesis and pedagogy that Socrates presents throughout the *Republic*.

References

Benardete, Seth. *Socrates' Second Sailing: On Plato's Republic.* Pbk. ed. 1992. Chicago: University of Chicago Press, 1989.

Brann, Eva T. H. *The Music of the Republic: Essays on Socrates' Conversations and Plato's Writings*. 1st Paul Dry Books ed. Philadelphia: Paul Dry Books, 2004.

Dobbs, Darrell. "The Piety of Thought in Plato's Republic, Book 1." *American Political Science* Review 88, no. 03 (September 1994): 668–83. https://doi.org/10/drws6h.

Gifford, Mark. "Dramatic Dialectic in Republic Book 1." *Oxford Studies in Ancient Philosophy* Xx, no. Summer 2001 (2001).

Jackson, B. Darrell. "The Prayers of Socrates." *Phronesis* 16, no. 1 (1971): 14–37. http://www.jstor.org/stable/4181854.

Nightingale, Andrea. "The Philosopher at the Festival: Plato's Transformation of Traditional Theoria." In *Seeing the Gods: Pilgrimage in Greco-Roman and Early Christian Antiquity*, edited by Jas' Elsner and Ian Rutherford, n.d.

Plato. *The Republic of Plato*. Translated by Allan Bloom. 2nd ed. New York: Basic Books, 1991.

Rosen, Stanley. *Plato's Republic: A Study*. New Haven, Conn. London: Yale University Press, 2008.

Rudebusch, George. "Dramatic Prefiguration in Plato's Republic." *Philosophy and Literature* 26, no. 1 (2002): 75–83. https://doi.org/10.1353/phl.2002.0017.

Segal, Charles. "'The Myth Was Saved': Reflections on Homer and the Mythology of Plato's Republic," 1978, 315–36. http://www.jstor.org/stable/4476064.

Strauss, Leo. *The City and Man*. 6. Dr. Chicago: Univ. of Chicago Pr, 1997.

A New Distinction in Meta-Ethics

David DeMatteo

Introduction

The purpose of this paper is to make a new distinction in meta-ethics. Specifically, I will distinguish between externalism and internalism about normative principle validity (hereafter EINP). Given that the whole internalism/externalism schema has been applied to matters as diverse as mental content, epistemological justification, moral judgment, and reasons[1], it might seem as if there's no need to make further use of what is already becoming an overworn trope. But in this essay, I'll argue that my new employment of the distinction is not only conceptually original, in the sense that it isn't reducible to any other uses of the distinction, but also useful for clarifying our thought about normative principles. By normative principle, I mean some proposition that states that we have reasons to act in some particular manner. For example, the hypothetical imperative "You have reason to do whatever fulfills your ends" is a normative principle, since it provides us with reasons given that we have certain ends. "You should help others in need" is also a normative principle, even if it doesn't ex-

plicitly include the language of reasons, since it tells us that in certain situations we have reason to engage in certain types of actions. Normative validity, meanwhile, refers to whether a principle serves as a normative guide on our conduct by generating reasons which we ought to consider in action (Korsgaard 2008, 31, Korsgaard 2014, 80). This is different from a more robustly realist conception of validity which holds that some principle can be valid regardless of whether it generates reasons which we ought to consider in action. Principles, in this picture, are always normatively valid for some agent or set of agents, and what we will ultimately be searching for in this paper is what the conditions are for a given normative principle to be valid for some agent. A reason, meanwhile, is simply a consideration which counts in favor of acting in a certain manner (Nagel 2008). Note that these reasons don't have to be decisive: they might be outweighed by other reasons.

By making this new distinction, I mean to illuminate a certain conceptual space in which one can stake out various claims. When philosophers distinguish, say, between mental content externalism and internalism, they mean to point to a spectrum of various positions that can be taken about how the contents of our mental acts are constituted. Similarly, by employing EINP, I will attempt to show how a range of philosophical theories can be mapped along two opposing poles. This will only

be a useful device, of course, if the distinction I'm making is conceptually novel (I.e, picks out a real distinction) and is hence irreducible to a variety of other schemas which aim to accomplish a similar task. Therefore, the first portion of this paper will be dedicated to providing an exposition of EINP and show that it isn't reducible to reasons internalism/externalism, moral judgment internalism/externalism, and realism and anti-realism about ethical propositions. With the basic conceptual originality of the distinction in place, I'll then advocate one particular position along the principle internalism/externalism spectrum, which I call the convergence position. Finally, I'll conclude by making some closing remarks on the philosophical worth of the distinction by both situating it within a broadly Kantian problematic and showing how it can help resolve certain disputes in metaethics.

The Distinction

Are normative principles valid for some agent because of beliefs, dispositions, and attitudes that agents have, or are they valid because of certain facts about the world? This is the basic issue around which EINP revolves. To the close reader, it might seem like this question creates a dichotomy where none exists: aren't facts about agents also facts about the world? We should distinguish agent-neutral facts and agent-relative facts. Agent-neutral facts are those which pertain

to all agents, like, say, the fact that they are agents and have desires. Agent-relative facts are those which are only true of certain agents, like the fact that I have a desire to someday enter the philosophy profession. So now we can clarify our initial question: are normative principles valid for an agent because of particular beliefs, dispositions, and attitudes that individual agents have, or are they valid because of certain facts about the world and agent-neutral facts about agents? Some examples are probably in order. Agent-neutral facts are those like the fact that we have desires and are capable of deliberating about what we ought to do. Kantian moral theory holds that general facts about practical reason and the nature of agency can be used to deduce principles which are binding on agents (or normatively valid). Notably, for Kantian moral theory, it is an agent-neutral fact that to act is always to act according to some maxim. This fact about the nature of agency is used by the Kantian to argue for the necessity of universal law (see Korsgaard 1996). Agent-relative facts are those like the fact that I have a desire to help others. On certain self-interest theories of a Humean variety, a principle claiming I ought to help others would not be normatively valid for me unless I already possessed such a desire.

By saying that principles are valid for some agent because of either particular beliefs and dispositions or

agent-neutral facts, I don't mean to imply that there is some causation going on here. "Because" merely indicates that there's some relationship of dependence: in an externalist picture, a principle might not be valid if certain facts about the world are not true, but that doesn't entail that the truth of those facts cause the principle to be valid. They are rather conditions for its normative validity, and our inquiry here is really into what the conditions are for any given normative principle's validity. Are these conditions basically bound up with the particular attitudes of agents, or are they dependent on more general facts about the world?

Internalist positions hold that the normative validity of principles is dependent in various respects on the propositional attitudes of particular agents. An extreme internalism about principles will thus hold that a principle is valid for an agent by the mere fact that an agent regards it as being valid. A more moderate internalist will assert that there is a complex web of beliefs which are necessary for a principle to be valid, but it is still ultimately a matter of an individual's attitudes and beliefs. In this case, the agent might need to have certain beliefs in not only the validity of the principle, but also other propositions which are rationally entailed by the principle. Or they might have to simply rationally believe in the principle's validity, and not hold that it is valid merely because of some personal idiosyncrasy.

That sets up constraints on the internal configuration of beliefs which can vouchsafe a principle's validity, but the validity of those principles is still wholly dependent on agent-relative facts.

By contrast to these internalist positions, an extreme externalism claims that a normative principle's validity does not depend on the beliefs and attitudes any given agent has, and can be valid for a particular agent even if it is impossible for that agent to rationally consider it as valid. The only criterion for its validity is that the facts about the world which are required for the principle's validity be true. There is also a weak externalist position which is possible to stake out in this conceptual space, which holds merely that the current configurations of beliefs and desires that any agent has cannot on their own be a sufficient condition of the principle being valid for that same agent. In other words, some agent-neutral fact (or fact about the world) must be true for the principle to be valid. Corresponding to this form of weak externalism is a weak internalism, which asserts that subjects must be at least capable of rationally regarding the given practical principle as being normatively valid, which would also entail, of course, holding any beliefs which are entailed by the principle. Note that weak internalism and externalism are not exclusive, but rather eminently compatible with one another. We'll call the fusion of these two positions the convergence po-

sition. Later, I'll provide a limited argument in defense of it. For now, though, we need to defend EINP itself.

A Defense of the Distinction

The general distinction between internalism and externalism has occasioned fierce debates in a wide swathe of philosophical sub-fields. As far as I can tell, though, nobody has yet applied it to practical principles themselves, and so it's worth clarifying why the distinction made here is genuinely different from 1) judgement internalism/externalism, 2) reasons internalism/externalism, and 3) realism and anti-realism about normative claims. Judgement internalism/externalism concerns whether moral judgments necessarily provide us with motivations for action.[2] A judgement internalist holds that whenever I make some moral judgement, it will provide me with a motivation for action. If I decide that it is wrong to consume animals, I will therefore have a motivation to stop consuming animals. Judgement externalists hold the opposite position: it's possible for me to judge that meat-eating is immoral without having any motivation whatsoever to cease eating meat. I'll argue that EINP has a different domain from judgment internalism/externalism by demonstrating that the two distinctions don't neatly map onto one another. A philosopher could easily hold, for example, both moderate internalism about moral principles and extreme externalism about moral judgements. If this is the case, the

validity of our moral principles will ultimately be tethered to the particular desires of agents, but the moral judgements we make using those principles won't ever provide us with motivation on their own. This is quite possible, precisely because what makes a normative principle valid is quite different from how that principle serves to both motivate moral judgements and how those judgements in turn interact with our particular attitudes. We can also flip this around, and consider a philosopher who holds weak externalism about principles and moral judgment internalism (A given moral judgment necessarily provides a motivation for the agent to act on it). In fact, this is arguably the position of Thomas Nagel in *The Possibility of Altruism* (2008), who combines a robust moral realism rooted in the external objectivity of moral principles based on agent-neutral facts with an internalism about moral judgments. This is a tempting position for any moral realist who wants to hold that morality's principles are unconditionally binding on individuals due to non-agent relative facts, and that individual moral judgments provide agents with the motivations to follow them.[3]

The debate about reasons' internalism and externalism has a closer connection to practical principles, but, as we'll see, the two distinctions still legislate over different domains.[4] The crux of the difference between internalists and externalists about reasons concerns the

relation between reasons and motivational facts.[5] Internalists hold that for something to be a reason for an agent, it must be tied in some way to their motives and desires (Arkonovich 2013, 210). Externalists hold the opposite: something can be a reason for an agent even it doesn't bear any relation to their propositional attitudes. A reasons internalist will argue that a maddened serial killer has no reason to cease his murders if there is not any connection between this reason and their present motives and desires. An externalist will hold the opposite: there's reason for them to cease their murders even if there is no deliberative route which might connect their current desires and motives to the moral reason to cease killing. Externalists don't have to hold that all reasons are external – they're merely committed to the proposition that some reasons are not internal.

As we'll see, it's possible to hold an internalist or externalist position in one domain without doing so in another. The instrumental principle is illustrative here. Roughly, the instrumental principle states that if we have an end, we have some reason to pursue the means to that end. Now note that I might be some form of an externalist about the instrumental principle (there must be some non-agent relative facts that the instrumental principle depends on to be valid), while also cleaving to an internalist account of instrumental reasons (instrumental reasons must be capable of motivating an agent

for them to be reasons). This account of the instrumental principle is actually quite appealing, because it explains why it is binding on all agents while simultaneously providing reasons that always must be capable of motivating an agent! Any account of the instrumental principle that made its reasons "external" and thus incapable of always potentially motivating agents would be quite bizarre – but we also want to understand the principle itself as having a sort of validity which is due to facts that don't just pertain to particular agents (Bedke 2009, Jollimore 2005). We can flip this around in the case of moral principles, and imagine a case where I am a weak internalist about some moral principle, but also an externalist about certain reasons. In this case, I will believe that this moral principle's validity must be capable of being recognized by an agent for it to be valid for them, but also think that certain reasons might bind those agents regardless of the desires and attitudes which they might have. That is eminently sensible, and in fact might be the best account of moral principles there is: We might want to say that someone has reasons to respond to some principle even if they had no way of being motivated by them, but we also might be doubtful that a principle could be valid for them if they had no way of recognizing it as valid.

Think, for example, of the moral principle: "Whenever it is not unnecessarily burdensome for you, help others".

Now suppose that for some agent, they have no desire or set of motivations to actually aid fellow human beings. In this case, we might say that they have an external reason to follow the normative principle in question, but that the reasons generated by the principle itself are internal because they are only reasons for the agent if they are capable of recognizing the principle's validity. We would thus be externalists because about reasons because we hold that there are some external reasons but also be internalists about principle validity.

Once we approach extreme internalism and externalism about principles, matters do get somewhat muddier. It would be very odd to be, say, an extreme internalist about moral principles and also be an externalist about the reasons that those moral principles give us. If all that is necessary for a principle to be valid is that an agent regard it as valid, then it would be difficult to understand how the reasons that such a principle provides could be externally binding on individuals regardless of their attitudes. It is also difficult to combine an extreme externalism about moral principles (the validity of a moral principle does not depend on any attitude, motivation, or belief an agent has, and would be true even if they were not capable of regarding it as valid) with reasons internalism. Presumably, if the validity of a moral principle didn't depend on any agent-relative facts, then the reasons it provides wouldn't either. Therefore, there are

some relations of entailment between these two distinctions. But this doesn't mean that the two distinctions aren't, in fact, distinct - especially because the entailment relations are incomplete.

Lastly, it's worth saying something about why EINP isn't simply reducible to realism and anti-realism about normative facts. There are three responses we can give, based on three renderings of what it means to be a "moral realist". First, suppose we treat moral realism and anti-realism as mapping onto principle validity itself. In this picture, normative realists are those who hold that normative principles can give reasons, and anti-realists are those who argue the opposite. In this case, we can distinguish between the two positions by noting that EINP assumes realism about normative validity, and then inquires into the conditions for a given principle to have that validity. If we treat moral realism and anti-realism as defined by their positions on the mind-dependence of moral facts, then we can show that it is possible to be both a moral realist and principle validity internalist. One might hold, for example, that moral principles exist in some sense independently of us, but only have normative validity if various conditions hold true which are agent-relative. Christine Korsgaard has pointed out that even if moral facts exist, they need some way of "getting a grip" on us (Korsgaard 2014). In other words, even if there exist moral principles irrespective of us,

that would not imply that those principles are necessarily normatively valid for us. And it might just be that for those moral principles to be normatively valid, they need certain agent-relative facts to also be true. This opens up the path to a fairly deep rift between moral realism and principle externalism. Conversely, one might take the position that it is possible for moral principles to be completely dependent on various facts about agents and have no existence apart from agents, but have those facts be solely agent-neutral ones. In this way, moral principles would be "mind-dependent" in the sense that they are constituted by various truths about agents, but their normativity would still be "externalist" insofar as they are not dependent on agent-relative facts. This would effectively combine moral anti-realism (at least if anti-realism is understood asserting that moral propositions are mind-dependent) and principle validity externalism. Lastly, we can think of moral realism as involving two claims: the capacity of moral judgments to be true or false, and the truth of most ordinary moral judgments (See Sayre-McCord 1986, from Joyce 2015). Just as internalism and externalism about principle validity presupposed the validity of principles, so it presupposes the capacity for moral judgments to be true. Consider if some normative principle was untrue: in that case, it would not provide us with reasons for acting. It would thus be of no use inquiring into its validity, since it would have no validity. If we are going

to inquire into the conditions for the normative validity of some principle, we thus must assume realism. This parallels the earlier case, where the normative validity of a principle was itself a condition for EINP rather than a distinction that could be equated with it.

A Defense of the Convergence Position

Earlier, I detailed what I called the convergence position. Recall that this position holds both A) that agents must be capable of rationally regarding some principle as being valid, which means holding various beliefs that a principle entails and B) that there must be some facts which are not agent-relative if the principle in question is to be true. Note that this is merely a thesis about the necessary conditions of principle validity – it does not state what facts are sufficient for a principle's normative validity, which is a much trickier task (and one that I will refrain from in this paper).

Let's first consider A). This basically amounts to the thesis that if a particular agent has no way of rationally regarding a principle as being valid, it can't be normatively valid for them. The negation of this position is that a principle could be normatively valid for an agent even if they can't rationally regard it as being valid. This is, in effect, an extreme form of externalism, which can be dismissed solely on the basis of the principle of ought implies can (hereafter designated OIC). Following Johnny Anomaly's article, OIC claims that no act "can be morally required if it is beyond human capacities to perform." (Anomaly 2008) Reformulated to apply to principles, OIC claims that no principle can be normatively valid if it is impossible for human beings to rationally regard the principle as valid. In other words, if we cannot regard it as valid, it cannot be valid, since it could never be normative for us. For a principle to be normatively valid, it must be capable of rationally guiding human action - but if we cannot rationally accept the grounds on which that principle rests, then the principle will not be capable of rationally guiding our action.

Now let's consider B). This is the claim that if some given principle is to be true, there must be some non-agent-relative facts that are true. Its converse is that there do not need to be any agent-relative facts that are true for a principle to be normatively valid. There are a few ways to stake out this position: one might, firstly, be an extreme internalist who holds that a principle could be made true solely by an agent holding it to be true. In this case, the agent in question doesn't need to have any other beliefs or attitudes except one endorsing the truth of the principle. This would entail the rather bizarre position, though, that all sorts of principles are normatively valid, such as "whenever I can, I will harm myself and others" - even if the agent also believes that pain is objectively wrong. To avoid conclu-

sions like this, one might endorse a moderate internalism, which requires an agent to have a sort of rational consistency in their beliefs that requires believing in any propositions that follow from the validity of a principle. Let us call these beliefs "presuppositions". Take, for example, the prudential principle, which states that one has some reason to do what is in one's future self-interest. Suppose that this principle "presupposes" the existence of a continuously existing self. Now assume that such a self can be demonstratively shown to not exist, and that Johnathan labors under the delusion that it does. Is the principle of prudence normatively valid for Johnathan? It makes most sense to say that he might regard this principle as normatively valid, but that it lacks real normative force because it is based on false assumptions. Another example might make this more lucid. Suppose that moral principles presuppose the existence of other agents who feel pain, but there are in fact, no agents like that which exist, and they in fact cannot exist. Johnathan lives in a solipsist world, where his peers are all elaborate automatons. But he mistakenly believes them to be real persons. Is the moral principle truly normatively valid in this case? Johnathan might believe it to be normatively valid, but surely it is not a principle that he ought to act on, all-things-considered. The principle is only mistakenly believed to be normatively valid.

The above cases demonstrate that the various presuppositions of a principle must also be true if a principle is to be normatively valid. If those presuppositions assert the truth of agent-neutral facts, as they likely will, then the form of weak externalism we've already discussed will be true, since the given principle's validity will depend on various agent-neutral facts. This establishes both the clauses of the convergence position, and thus demonstrates that it is true.

The Philosophical Merit of the Distinction

How tethered must morality be to human life and human practices? This distinction basically speaks to that question. Rather than asking about reasons or moral motivations, however, it inquires into the validity of normative principles themselves. Can a normative principle be normative for some human being if they found themselves incapable of recognizing its validity? Are these principles rooted in deep facts about the world, or merely in the particular desires of agents? These are the questions that this distinction grapples with.

It is concerned with what is basically a Kantian problematic: what are the conditions for the validity of various ethical principles? In answering this question, one must make a transcendental argument[6] which moves from the normative validity of a given principle to the conditions for the validity of that same principle. The

distinction itself thus encourages moral philosophers to make use of transcendental argumentation, which in recent years has made something of a come-back in analytic ethics with the work of Korsgaard (*Creating the Kingdom of Ends* 1996). And of course, we can ask whether the various conditions discovered by transcendental argumentation are rooted in 1) the idiosyncrasies of various subjects, 2) the structure of the world subjects live in, or 3) the *a priori* conditions of agency itself. Notably, those who take these last two positions are both put in the externalist camp, but they might have quite different philosophical predilections: whereas many of the former consider themselves realists who believe moral facts exist on the same plane as platonic mathematical entities (See Parfit 2011), those in the latter group often believe that moral principles are valid because of certain facts about the structure of rational agency (Korsgaard 1996).

If this distinction has merit, it will not just be, though, in its responsiveness to perennial philosophical questions. Its value must also lie in its aim to clearly explicate how a variety of meta-ethical stances can be mapped, and to clarify our philosophical discourse in doing so. For example, there has long been a dispute over Korsgaard's claim that normative claims need some way to "get a grip" on us (Korsgaard 1996). Yet this distinction makes clear that this is a different manner from the

whole question of whether normative principles exist independently from us. One might be a realist about the ontological status of normative principles but an internalist about their normative status. That is, normative principles might very well be the sort of thing that require some connection to our propositional attitudes to have validity, but also exist independently of agents. EINP thus opens up the philosophical space to notice that ontological status and normative status are indeed distinct, and thereby enables us to increase the sophistication of our philosophical discourse. This might be invaluable to philosophers who wish to hold that various sorts of normative principles do indeed possess a sort of mind-independence, but who also don't want to forsake their connection on a normative level to motivational states. At its best, then, this distinction might serve as a heuristic tool, enabling philosophers to more thoroughly clarify where they stand. And insofar we conceive of philosophy as itself a practice of conceptual clarification, the art of making such distinctions is not merely an aide to the practice of philosophy, but itself a form of philosophical practice.

Notes

1. For a paper that nicely reviews and criticizes externalist accounts of mental content, see (Farkas 2008). For a review of the literature around epistemic justification, see the bibliography of (Bonjour and Sosa 2008).

2. In this I am following Russ Shafer-Landau's early definition of moral internalism in "A Defense of Motivational Externalism" (2000: 270). Shafer-Landau also offers an overview of the literature around moral internalism and externalism.
3. Of course, there is an anti-realist argument that depends on motivational judgment internalism, outlined on page 121 Shafer-Landau's 2003 book on moral realism. It moves from an acceptance of motivational judgment internalism and motivational humeanism (beliefs do not yield motivational states) to a robust anti-realism. The key fact here is that motivational judgment internalism alone does not produce an anti-realist position. And indeed, several philosophers who are moral realists of various stripes have rejected motivational Humeanism but accepted judgment internalism, like McDowell (1978) and Scanlon (2000).
4. For some defenses of reasons internalism, see Williams *Ethics and the Limitations of Philosophy* (1985), and Cowley "A New Defense of William's Reasons-Internalism" (2005). For an overview of internalist positions, see Arkonovitch's "Varieties of Reasons/Motives Internalism" (2013). For some broad-sides against internalism, see Setiya's "Against Internalism" (2004), Parfit's "Reasons and Motivation" (1997) , and Brewer's "The Real Problem with Internalism about Reasons" (2002).
5. I take this definition from (Finlay and Schroeder 2017).
6. I.e, an argument that moves from the validity of some given principle or practice (experience, knowledge, etc.) to the conditions for the possibility of its validity.

References

Anomaly, Johnny. 2008. "Internal Reasons and the Ought-Implies-Can Principle." *The Philosophical Forum* 469-483.

Arkonovich, Steven. 2013. "Varieties of Reasons/Motives Internalism." *Philosophy Compass* 210-219.

Bedke, Mathew. 2009. "The Iffiest Oughts: A Guide of Reasons Account of End-Given Conditionals." *Ethics* 672-698.

Brewer, Talbor. 2002. "The Realm Problem with Internalism about Reasons." *Canadian Journal of Philosophy* 443-473.

Copp, David. 2007. *Morality in a Natural World: Selected Essays in Metaethics*. Cambridge University Press.

Cowley, Christopher. 2005. "A New Defence of Williams' Reasons-Internalism." *Philosophical Investigations* 346-368.

Farkas, Katalin. 2003. "What Is Externalism?" *Philosophical Studies* 187-208.

Finlay, Stephen, and Mark Schroeder. 2017. *Reasons for Action: Internal vs. External.* August 18. Accessed November 22, 2018. plato.stanford.edu/entries/reasons-internal-external/.

Jollimore, Troy. 2005. "Why is Instrumental Rationality Rational? ." *Canadian Journal of Philosophy* 289-307.

Korsgaard, Christine. 1996. *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press.

Korsgaard, Christine. 1996. "Kant's Formula of Humanity." In *Creating the Kingdom of Ends*, by Christine Korsgaard, 106132. Cambridge University Press.

Korsgaard, Christine. 2008. "The Normativity of Instrumental Reason." In T*he Constitution of Agency: Essays on Practical Reason and Moral Psychology*, by Christine Korsgaard. Oxford University Press.

—. 2014. The Sources of Normativity. Cambridge University Press.
McDowell, John. 1978. "Are Moral Requirements Hypothetical Imperatives?" *Aristotelian Society* Supplementary volume 13-42.
Nagel, Thomas. 2008. The Possibility of Altruism. Princeton University Press.

Parfit, Derek. 2011. *On What Matters.* Oxford University Press.
Sayre-McCord. 1986. "The Many Moral Realisms." Southern Journal of Philosophy 1-22.

Sayre-McCord, Geoffrey. 2007. *Essays on Moral Realism*. Cornell University Press.

Scanlon, Thomas. 2000. *What We Owe to Each Other.* The Belknap Press.

Setiya, Kieran. 2004. "Against Internalism." *Nous* 266-298.

Shafer-Landau, Russ. 2000. "A Defense of Motivational Externalism." *Philosophical Studies* 267-291.

—. 2003. *Moral Realism: a Defence*. Oxford: Oxford University Press.

Williams, Bernard. 1985. *Ethics and the Limitations of Philosophy.* Harvard University Press.

Digging Beneath Wittgenstein's Bedrock: An Attempt to Specify What is Shared in a Common Form of Life

Jonah Goldberg

Introduction

In his later work, *Philosophical Investigations* (PI), Ludwig Wittgenstein points out a number of problems with the notion of rule-following, noting specifically the difficulty of identifying how it is we know what to do (or, indeed, even what it is we are doing) when we follow a rule.[1] He observes that we have a tendency to locate the directing power of a rule in an interpretation of the rule without noticing that the interpretation is just another rule whose directing power still seems to depend on being further interpreted, *ad infinitum*. Wittgenstein thinks that this tendency arises from a mistaken belief that "every action according to a rule is an interpretation," when in fact, "there is a way of grasping a rule which is not an interpretation, but which, from case to case of application, is exhibited in what we call 'following the rule' and 'going against it.'"[2] When asked by the text's interlocutor to specify what this "way of grasping a rule which is not an interpretation" is, however, Wittgenstein's response is unsatisfying. He writes, "If this is not a question about causes, then it is about the justification for my acting in this way in complying with the rule. Once I have exhausted the justifications, I have reached bedrock, and my spade is turned. Then I am inclined to say: 'This is simply what I do.'"[3] By this, Wittgenstein means that the answer to the question being asked lies beyond a point past which no further analysis can be conducted. That is the sense in which Wittgenstein believes he has reached bedrock.

This remark comes at an awkward point in Wittgenstein's argument. He has demonstrated that the use of formal systems (including language and mathematics) is in some sense strictly underdetermined by the rules that constitute those systems. This leaves it deeply unclear how people can communicate with one another using language at all because it seems as if each individual should not be able to predict how any other individual will interpret and apply any given rule, including, for example, the definitions of words. And yet, evidently, this is not a problem people have. Wittgenstein, then, needs to provide some explanation of what it is that allows that us to communicate with one another in spite of his rule-following considerations. He acknowledges this (obliquely) but pronounces any analysis of what it is that enables our communication to be beneath "bedrock."[4] He labels this feature we cannot analyze that

makes communication possible a shared "form of life."[5] It's clear in the PI that Wittgenstein envisioned forms of life as being in some way related to customs, conventions, and social practices, but his conception of the precise relation between forms of life and customs seems blurry, plagued by, among other things, a profound ambiguity about where exactly bedrock begins.[6]

In this paper, I will attempt to clarify these ideas. I will use the hints provided in the PI regarding the nature of forms of life to defend the idea that a shared form of life is a shared conception of the terms of a "language-game"[7] (a use pattern in language) and that this conclusion does not lie beneath bedrock. Rather, the bedrock begins somewhere shortly beneath it. I will then explain how this allows us to specify what precisely Wittgenstein means by his claim that meaning is use.[8] Finally, I will defend this position against the accusation that it rests on a violation of Wittgenstein's metaphilosophical commitments, as has been suggested of similar ideas in the secondary literature on the PI.

Communication and Locating Meaning

The basic Wittgensteinian problem is, at its core, the observation that a degree of ambiguity persists in every specification of a rule, and that consequently, no specification of a rule can adequately explain what is to follow or to violate that rule. For example, when we

observe others adding two numbers together, it's impossible for us to be sure whether the operation that they performed was identical to the operation that we perform when we do addition. They may have reached the same sum that we would have, but they might have reached it by calculating √(X + Y)2 instead of merely X + Y. Similarly, when we do addition, it's impossible for us to know that the operation that we're performing is, fundamentally, what addition is. This is because even if we read in our math textbook that addition can be fully described in terms of the successor function, a counting algorithm, and we simply perform deconstructed addition using that counting algorithm, we are still employing tacit rules to apply that counting algorithm to the particular sum in question. We are using a translation rule to understand addition in terms of the counting algorithm, and we are using a succession rule to use the counting algorithm itself. To the extent that addition is indeterminate, any decomposition, analysis, or interpretation of addition (or of addition's components) will be similarly indeterminate, *ad infinitum*. Applying this to language, Wittgenstein contends that the meanings of words are really just rules and are thus subject to rule-following paradoxes. For instance, when someone says, "Inside of eggs, there is yolk," she might mean to define yolk only as "that which is inside of an egg" rather than as the yellow stuff that we conventionally understand to be yolk, which is absent from some

eggs.[9] When cracking eggs that lack this yellow stuff, we would disagree with our interlocutor about whether or not they contain yolk. In such cases, we cannot properly say that our interlocutor is wrong. At most, we can say that her use is atypical.[10]

This observation makes it unclear why we are able to communicate with one another with so little difficulty. In explanation, Wittgenstein suggests that interpersonal communication is made possible by a shared "form of life," that this is what enables us to predict how other people will interpret and apply rules in language so that the words we say can have public meaning.[11] Accordingly, that which is shared within a common form of life must be the source of the demands that rules seem to place on us with respect to how they should be followed, for it is by referring to these demands that we predict how others will follow rules. We assume that they feel the normative push of the same demands that we do. In cases when this assumption is correct, communication is possible. Therefore, speech has meaning when (and because) the demands the speaker felt (the applicable) definitional rules placed on her in speaking are similarly felt by the listener in listening. The demands must be, so to speak, mutually legible. If to understand what someone is saying, one must understand the demands she feels the rules of her speech place on her, then the status of demands in speech is consistent with what we

conventionally call "meaning." One understands what someone says when one understands the demands in her language; one understands what someone says when one understands the meaning of her language. Meaning, like these demands, is that which is understood in language that is understood.

As noted above, the source of the demands that a particular set of rules seems to place on us with respect to their execution must be the thing that is shared in a common form of life. The question, then, is: What is that source? We have already seen how that source cannot be mere interpretations of the rules at hand, for in an interpretation, "one expression of a rule is substituted for another."[12] Maybe, however, these demands could emerge from a mental state, like, for example, a person's intentions. Wittgenstein, however, disagrees. The most obvious problem with this view is that mental states are not publicly observable. This, on face, makes them a difficult place to locate the source of meaning because it is so easy to misidentify them in others. Some people betray little of what they are thinking on their faces; others engage in deliberate deception about the content of their thoughts. Surely it cannot be the case that the factor by virtue of which we understand each other's language lies locked inside our heads, hidden entirely from others' view.[13]

The larger problem with locating the source of meaning among mental states, however, is that for Wittgenstein, it's not clear that even we ourselves can correctly identify our own mental states. Part of the problem here relates to Wittgenstein's doubts about whether our concepts of mental states are consistently and coherently specified in the first place. He offers the following analogy to illuminate this concern:

Suppose that everyone had a box with something in it which we call a "beetle." No one can ever look into anyone else's box, and everyone says he knows what a beetle is only by looking at his beetle. – Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. – But what if these people's word "beetle" had a use nonetheless? – If so, it would not be as the name of a thing. The thing in the box doesn't belong to the language-game at all; not even as a Something: for the box might even be empty. – No, one can "divide through" by the thing in the box; it cancels out, whatever it is.[14] In short, Wittgenstein believed that our mental states are subject to the exact same problem of underdetermination as any other lens through which one might offer an analysis of a rule because their private content can neither clarify nor justify the public meanings of our spoken words. The fact that we feel we know what we mean when we speak offers no es-

cape from his rule-following paradoxes. He even goes so far as to claim, analogically, "If God had looked into our minds, he would not have been able to see there whom we were speaking of."[15]

The Private Language Argument

Wittgenstein provides a systematic analysis of why this is the case in his Private Language Argument (PLA). In the PLA, Wittgenstein argues that private ostensive definitions are impossible, so the definitions of words must begin and end in their public use.[16] The argument begins by proposing the notion of a private language, a language that, in principle, could be understood by just one individual and no others.[17] The conclusion of the private language argument is that such a language is inconceivable. The crux of Wittgenstein's thought on this topic revolves around the notion of private meaning. Wittgenstein viewed private meaning as nonsensical in concept, positing instead that meaning must necessarily arise from public criteria of justification. Accordingly, words that, for one reason or another, cannot possess public criteria of justification cannot mean anything. Wittgenstein offers the example of a person who decides to name a particular recurring sensation. He writes, "Let us imagine the following case. I want to keep a diary about the recurrence of a certain sensation. To this end I associate it with the sign 'S' and write this sign in a calendar for every day on which I have the

sensation — I first want to observe that a definition of the sign cannot be formulated."[18] At the simplest level, the reason the individual in question fails to define S as the sensation at hand is because merely associating the symbol S with this sensation repeatedly over time fails to produce any means of justifying whether any particular experience of a sensation is a case of S. As Wittgenstein puts it:

> Let us imagine a table, something like a dictionary, that exists only in our imagination. A dictionary can be used to justify the translation of a word X by a word Y. But are we also to call it a justification if such a table is to be looked up only in the imagination? — "Well, yes; then it is a subjective justification." — But justification consists in appealing to an independent authority — "But surely I can appeal from one memory to another. For example, I don't know if I have remembered the time of departure of a train correctly, and to check it I call to mind how a page of the timetable looked. Isn't this the same sort of case?" — No; for this procedure must now actually call forth the correct memory. If the mental image of the timetable could not itself be tested for correctness, how could it confirm the correctness of the first memory? (As if someone were to buy several copies of the morning paper to assure himself that what it said was true.)[19]

Wittgenstein's point is that attempting to confirm that a sensation is S by comparing it to one's memories of

previous times one wrote down S in one's calendar is mere "ceremony," as it requires one's memories of previous cases of S to be true cases of S and thus valid bases for comparison.[20]

The problem with this is not merely that one's memory may be unreliable. Let's say that one defined S for the first time (establishing its true definition) at 2:00 pm, and now, at 2:10 pm, one is experiencing another potential case of S. Surely, in this instance, one can trust one's memory of the base case. However, even in this instance, one cannot be justified in writing down S in one's calendar, for even in this instance, one has no means of specifying what about one's present sensation makes it a true case of S. By virtue of what quality or characteristic does it acquire its S-ness? Clearly, it isn't identical to the base case sensation in every respect; at a bare minimum, they occurred ten minutes apart. How can one be sure that the rule by which one is defining S sensations does not entail that they take place only at 2:00 pm? One might insist that one did not intend to define S sensations to include only examples of S that take place at 2:00 pm, but what exactly would one mean by "intend" here? Perhaps, one would mean that one did not predict that all future cases of S would take place at 2:00 pm. This prediction, of course, could be wrong. If it were wrong, would we not say that taking place at 2:00 pm is a defining characteristic of S? If we

would not, what exactly would it be that would make a characteristic of S a defining characteristic? If a given sensation shared only half of its characteristics with the base case of S, but the characteristics it shared with the base case included all of the characteristics that we judge to be essential to the base case, would we be correct in calling it an S? If so, we must ask what it is that makes a given characteristic of S "essential" to it. The obvious answer is: "These are the characteristics by virtue of which S sensations are S sensations," in which case we must admit that the reason we feel we are justified in calling the sensation in question an S is because some of its characteristics are characteristics that make us feel justified in calling sensations S. The circularity here is clear. As Wittgenstein put it, "'I commit it to memory' can only mean: this process brings it about that I remember the connection correctly in the future. But in the present case, I have no criterion of correctness. One would like to say: whatever is going to seem correct to me is correct. And that only means that here we can't talk about 'correct'."[21]

It is in this sense that the PLA is a special case of the rule-following paradoxes described earlier in the PI.[22] In the case of a rule-following paradox involving public language, however, the paradox alone is manifestly insufficient to deprive the language of its meaning. We know this because we succeed in understanding

the speech of others all the time. It must be the case that our words mean something, and so there must be something by virtue of which others understand them.[23] To Wittgenstein, that "something" must be able to serve as a public criterion of justification, and it is because a hypothetical private language would (by definition) lack any such public criteria of justification that it could not exhibit meaning. The criteria of justification must be public in order to produce meaning (even meaning to oneself, if such a notion can be considered coherent) because only public criteria allow one's justification to appeal to something independent of that which is being justified.[24] Recursive appeals back to one's own feelings or judgments, as in the case of a private language, cannot produce justifications for one's own feelings or judgments.

Linguistic Communities and Language-Games

Because the source of meaning cannot be a mental state, whatever is shared in a common form of life must consist only of information that can be displayed publicly. Perhaps what is shared in a common form of life is something like a culture; maybe the information in question is information about a linguistic community. This seems plausible but ill-specified. To isolate exactly what within a linguistic community must be shared in a common form of life in order for that form of life to give rise to meaning, we can consider Wittgenstein's re-

mark: "If a lion could talk, we wouldn't be able to understand it."[25] The question of what it is that this lion lacks as a result of which we could not understand it gets right to the heart of the issue. To untangle this, imagine a linguistic community of humorous mathematicians in which two language-games comprised all communication: telling jokes and factoring polynomials. If you just dropped the lion into this linguistic community and let it watch the mathematicians go about their factoring, it's difficult to imagine that even a lion with human-level intelligence would be able to predict how the mathematicians would factor a given polynomial merely by watching them factor a few dozen. This seems like it would remain true even if the lion had spent a good bit of time in human linguistic communities before. Remember, Wittgenstein's lion can talk; we imagine that in its head, it has a complete English dictionary. The issue is not that of a conventional language barrier. The barrier lies in its grasp of the terms of the language-game.[26]

Now, imagine that the highly intelligent, English-speaking lion were also a mathematician with lots of experience factoring polynomials. If such a mathematician lion were to immigrate from a community of similar lions into this community of humorous, human mathematicians, it seems likely it would be able to correctly factor polynomials alongside the human mathematicians without much difficulty. After all, in Wittgenstein's words, "Mathematicians don't in general quarrel over the results of a calculation."[27] In fact, it seems as if it would be able to factor polynomials with them (and produce the same answers they did) even if it didn't speak English. It would still know the terms of the relevant language-game, and that's what matters. However, it also seems clear that this lion, for all its knowledge of the English language and all of its mathematical ability, would nonetheless be unable to understand the human mathematicians' jokes, much less come up with jokes itself that would make them laugh. This is the result of its failing to grasp the terms of the relevant language-game. If it entered their community after working for several years as a writer for the comedy show Saturday Night Live, on the other hand, we would not expect humor to cause it any difficulty.

It should be clear from this example that the information about a linguistic community that must be shared in a common form of life in order for that form of life to give rise to meaning consists only of the terms of the relevant language-games, not anything about the community itself or its culture, per se. The terms of the relevant language-games alone constitute the source of the demands that the rules at hand place on us with respect to their execution. What, then, are these terms? Obviously, it can't be the case that they are rules themselves, and it further can't be that they are interpreta-

tions of rules. If that were the case, then a shared form of life could not serve as a foundation for meaning because it would offer no resolution to the problem of the infinite regress of interpretations. Relatedly, Wittgenstein suggests that when we "grasp" a rule, when we follow a rule with confidence in our way of execution, we necessarily do so "without reasons," that we "['exhaust'] the justifications,"[28] at which point we are "inclined to say, 'This is simply what I do.'"[29] In light of that, we must understand the terms of language-games not as rules or interpretations of rules but as the living practices that constitute the language-game, such that to grasp a rule is nothing more or less than to do those living practices, to play the language-game.[30]

Reaching Bedrock

This intersects rather neatly with a theory of where bedrock begins. Bedrock must begin at the level of the living practices themselves. This is to say that the first concept set in the Late Wittgensteinian scheme of which we can offer no further specification or analysis than is present in its naming alone is that of the living practices that make up the terms of language-games. Concretely, we can offer no complete account of the living practices that make up any single language-game, and we can offer no adequate explanation of the role that any particular living practice plays in establishing a given language-game. The simplest reason why this is

where our analysis must stop is because the structure of living practices is not propositional; it is not the sort of thing that can be fully represented in our language, in the same way that the melody of a song cannot be fully represented with words alone. Importantly, this is a necessary characteristic of these living practices, for if they could be fully represented by words alone, then they could be interpreted as rules, which would render them inadequate for the purpose of grounding meaning. Instead, they allow us to escape from the "inclination to say: every action according to a rule is an interpretation" and see the "way of grasping a rule which is not an interpretation."[31] From this vantage point, it's clear why this, in particular, is where bedrock must begin, why it's impossible to dig any deeper. We know that articulable rules can't fully describe the terms of language-games because we have observed the persistence of rule-following paradoxes at every level of interpretation; the fact that they necessarily undermine every linguistic description or analysis of a language-game that one could possibly construct demonstrates that the terms of language-games cannot be fully represented in language. Furthermore, in Wittgenstein's words, "Language is itself the vehicle of thought;" what we cannot express in language, we cannot think either.[32]

This specification of the relation between forms of life and bedrock lends itself to an elegant interpretation of

Wittgenstein's claim that meaning is use.[33] He writes: "Here the term 'language-game' is meant to bring to prominence the fact that the speaking of language is part of an activity, or of a form of life."[34] This directly acknowledges that forms of life are constitutively similar to activities, which clearly allows for a model in which forms of life consist of the living practices that make up the terms of language-games. He then provides his readers with an especially clear example of just such a living practice, observing that it might "[come] naturally to a person to react to the gesture of pointing with the hand by looking in the direction from fingertip to wrist, rather than from wrist to fingertip."[35] "Pointing with the hand" is a language-game whose terms include the demand that the observer to shift her focus in the direction indicated by the fingertip.[36] This term is a "living practice" because the demand in question only comes to exist in the actual use of gesture. An image of a hand with a finger extended on its own (without any context) "seems dead […] In use it lives."[37] Wittgenstein puts it in similar terms, writing: "Only in the process of understanding does the order mean that we are to do THIS. The order – why, that is nothing but sounds, inkmarks."[38] These remarks establish quite clearly in what sense meaning is use. Use is the process of vivifying a sign such that it becomes the sort of living practice that can make demands on participants in a language-game. Meaning is use in the sense that meaning

emerges directly from that process of vivification.

Addressing Metaphilosophical Objections

Many scholars of Wittgenstein, it's worth noting, would reject any attempt to tie his ideas about forms of life, bedrock, and meaning as use together in this way. They would, in Crispin Wright's words, view it as dangerously close to an attempt to answer "the constitutive question," toward which "his […] philosophical method seems to be conditioned by a mistrust."[39] Or as John McDowell put it:

> If one reads Wittgenstein as offering a constructive philosophical account of how meaning and understanding are possible, appealing to human interactions conceived as describable in terms that do not presuppose meaning and understanding, one flies in the face of his explicit view that philosophy embodies no doctrine, no substantive claims. This view of philosophy is what Wright describes as quietism.[40]

The Wittgenstein presented in this paper is without question a slightly more revisionary Wittgenstein, a Wittgenstein with less of an aversion to "substantive claims," than the Wittgenstein McDowell and Wright believe wrote the PI, but the Wittgenstein that McDowell and Wright believe wrote the PI is a Wittgenstein engaged in a fundamentally flawed philosophical enterprise.

McDowell and Wright deny that in the PI, Wittgenstein is endeavoring to provide a theory of meaning and understanding, especially any such theory that attempts to provide a full, positive account of how meaning and understanding are possible. They are correct in observing that the PI does not endeavor to entirely explain how meaning comes to exist. Wittgenstein believed that "Explanations come to an end somewhere."[41] It is misguided, however, to suggest that the PI does not amount to a constructive theory of meaning. While it's certainly true that the theory it presents leaves open some questions about how, specifically, we come to learn a language and to acquire a form of life (addressing them only with underexplored allusions to "training"), it does offer a positive doctrine of what meaning is. Meaning in the PI is that by virtue of which language is understood and that which is understood in that language. It's the reason for the sense we get that we are following a rule correctly that comes embedded in the realization, "Now I can go on."[42]

McDowell and Wright deny that this kind of ontological specification of meaning emerges from the PI, except in the most limited, negative sense (i.e. the ontology of meaning can't consist in an Augustinian notion of correspondence). Instead, they would interpret Wittgenstein's positive adumbration of meaning as nothing

more than a conceptual analysis, an inquiry into how it is that we use the word meaning itself. Given an uncontroversial reading of Wittgenstein's views of meaning and use, however, it becomes clear that those views leave no room for a purely conceptual analysis of the word "meaning" without any ontological assumptions baked in. Accordingly, it seems unlikely that Wittgenstein intended to limit his treatment of the concept of meaning to this grammatical investigation alone (though remark 43 makes clear that he did intend to address the grammatical question).

In remark 43, when Wittgenstein does engage in a conceptual or grammatical analysis of meaning, he points out that what we are really looking for when we look for the meaning of a word is very often the way that word is being used in a given context. In this sense, on a purely grammatical level, meaning is use.[43] But how are we to understand use? What should we be looking for when we try to identify it? Luckily, precisely this substantive question is the focus of a good portion of the PI. Wittgenstein, as we know, tells us that we are not to look to the mental state of the speaker because the word's intended use is not our concern; instead, we are to analyze the context in which the word was used to identify its actual use. This actual use, we are told, relates specifically to the word's function in the conversation, to the communicative work it performed.[44] In order to un-

derstand the communicative work a word performed in a conversation, however, one needs to already have a theory of how communication is possible. For this reason, it's impossible to conduct a Wittgensteinian analysis of how a word is typically used without referring to at least a basic theory of how communication works. Without one, the notion of "use in context" itself is incoherent; use to what end? This seems like a very good reason to believe that Wittgenstein intended to provide something like a theory of communication in the PI, if only to enable grammatical analyses.

Conclusion

Wittgenstein spends much of the PI elaborating at length why it should be surprising to us that we can successfully communicate with one another through language in light of the illusory status of the meanings of definitional rules. And when he acknowledges the obvious truth that we can, in fact, communicate with one another through language, he offers little more than the assertion that the phenomenon is, at its deepest levels, inexplicable in concept as explanation. Upon closer examination of the PI, however, the echoes of a more comprehensive model of meaning emerge. Meaning is possible in the presence of agreement on the terms of the relevant language-games; Wittgenstein refers to this as agreement in "form of life."[45] The terms of language-games escape from the infinite regress of interpretation by being of nonpropositional form; rather, they are living practices. Accordingly, understanding consists of nothing more than playing the language-game at hand, and use involves the vivification of signs into those living practices. Though everything "beneath" this level of analysis must be bedrock because otherwise, the terms of language games would be subject to interpretation, the mere existence of bedrock is no sign of a failed explanatory enterprise. All explanations have to end somewhere, but there is virtue in pressing on as far as one possibly can.

Notes

1. "A has written down the numbers 1, 5, 11, 19, 29; at this point B says he knows how to go on. What happened here? Various things may have happened; for example, while A was slowly writing down one number after another, B was busy trying out various algebraic formulae on the numbers which had been written down. After A had written the number 19, B tried the formula an = n2 + n – 1; and the next number confirmed his supposition. […] Or again, B does not think of formulae. He watches, with a certain feeling of tension, how A writes his numbers down, while all sorts of vague thoughts float through his head. Finally, he asks himself, 'What is the series of differences?' He finds: 4, 6, 8, 10, and says: 'Now I can go on.' Or he watches and says, 'Yes I know that series' – and continues it just as would have done if A had written down the series 1, 3, 5, 7, 9. – Or he says nothing at all and simply continues the series. Perhaps he had what may be called the feeling 'That's easy!'" (Wittgenstein,

PI: Remark 151)
2. Wittgenstein, PI: Remark 201
3. Wittgenstein, PI: Remark 217
4. Ibid.
5. Wittgenstein, PI: Remark 23, 241
Wittgenstein, PPF (xi): Remarks 341-345
6. Wittgenstein, PI: Remark 198
7. See Wittgenstein, PI: Remarks 7 and 23 for characterizations of what constitutes a "language-game."
8. Wittgenstein, PI: Remarks 1, 43
9. Laskow 2016
10. See, e.g. Wittgenstein, PI: Remark 151, inter alia
11. Wittgenstein, PI: Remark 241
12. Wittgenstein, PI: Remark 201
13. Wittgenstein, PPF (xi): Remarks 328, 333, 336, 355, 357, 362
14. Wittgenstein, PI: Remark 293
15. Wittgenstein, PPF (xi): Remark 284
16. Wittgenstein, PI: Remarks 243, 258, 265
17. Wittgenstein, PI: Remark 243
18. Wittgenstein, PI: Remark 258
19. Wittgenstein, PI: Remark 265
20. Wittgenstein, PI: Remark 258
21. Wittgenstein, PI: Remark 258
22. Wittgenstein, PI: Remarks 138-242
23. Wittgenstein, PI: Remarks 197, 355
24. Wittgenstein, PI: Remark 265
25. Wittgenstein, PPF (xi): Remark 327
26. I believe that in positing a talking lion, Wittgenstein means

for his readers to contemplate an acknowledged impossibility. Of course, to Wittgenstein, it is a mistake to think any actual lion would employ human concepts. Even to the extent that lions use concepts to communicate amongst themselves, they have no use for any concepts other than their own (and could not acquire alien concepts, in any event). In this remark, however, Wittgenstein means to express more than just that corollary to his broader argument. He tells us his lion can talk. While it is unclear precisely how we ought to interpret its faculty of speech, I prefer the following. We imagine the lion with a list of English words and their definitions in its head (i.e. a dictionary, as in PI Remark 265). Nonetheless, it will never use those words appropriately.
27. See for reference: Wittgenstein, PPF (xi): Remarks 341-345
28.Wittgenstein, PI: Remark 211
29. Wittgenstein, PI: Remark 217
30. Wittgenstein, PI: Remark 23
31. Wittgenstein, PI: Remark 201
32. Wittgenstein, PI: Remark 329
33. Wittgenstein, PI: Remarks 1, 43
34. Wittgenstein, PI: Remark 23
35. Wittgenstein, PI: Remark 185
36. This description doesn't fully capture the "living" dimension of this exercise, but it's an adequate approximation for these purposes.
37. Wittgenstein, PI: Remark 432
38. Wittgenstein, PI: Remark 431
39. Wright 1989, 305
40. McDowell 1992, 51

41.Wittgenstein, PI: Remark 1
42. Wittgenstein, PI: Remark 151
43. Wittgenstein, PI: Remark 43
44. Wittgenstein, PI: Remarks 92, 525, 665
45. Wittgenstein, PI: Remark 241

References

Laskow, Sarah, "Found: An Egg with No Yolk," Atlas Obscura, December 21, 2016, https://www.atlasobscura.com/articles/found-an-egg-with-no-yolk (accessed December 2018).

McDowell, John. "Meaning and Intentionality in Wittgenstein's Later Philosophy." *Midwest Studies in Philosophy* 17, no. 1 (September 1992): 40–52. https://doi.org/10.1111/j.1475-4975.1992.tb00141.x.

Wittgenstein, Ludwig. "Philosophical Investigations," in *Philosophical Investigations*. Revised 4th ed. Translated by G.E.M. Anscombe. Edited by P.M.S. Hacker and Joachim Schulte. Malden, MA: Wiley-Blackwell, 2009. (PI)

Wittgenstein, Ludwig. "Philosophy of Psychology – A Fragment," in *Philosophical Investigations*. Revised 4th ed. Translated by G.E.M. Anscombe. Edited by P.M.S. Hacker and Joachim Schulte. Malden, MA: Wiley-Blackwell, 2009. (PPF)

Wright, Crispin. "Critical Notice." *Mind*, New Series, 98, no. 390 (1989): 289-305. http://www.jstor.org/stable/2255134.

Justifying Extraterritorial Political Obligations

Sun Woo Lee

Introduction

Selling or possessing marijuana is strictly prohibited in South Korea. If Kim, a Korean citizen took a short trip to California (where recreational marijuana is legal) and smoked weed there, South Korea still claims the legal authority to punish Kim. The Korean constitution claims authority to exercise extraterritorial jurisdiction over certain offenses not limited to drug use, soliciting prostitution, featuring in porn, and gambling. This means that Korean citizens are prohibited from engaging in those acts even outside Korea, notwithstanding the legality of those acts where they are conducted. The central question of this paper concerns whether imposing political obligations that apply to nationals beyond the bounds of its territory, as in the case of Kim, can be justified by liberal accounts of political obligation. I believe there isn't a clear answer to this question. This paper examines some paradigmatic accounts of political obligation and evaluates the reasons they can give for extraterritorial political obligations. Even when we assume their relative success in

accounting for political obligations to particular states in general, the paper argues that none of them are able to adequately account for political obligations that extend beyond its territory. Hence, it concludes that there isn't a satisfactory substantive, liberal justification for extraterritorial political obligations for the time being.

Extraterritorial Jurisdiction and Political Obligations

This paper is interested in justifications of extraterritorial political obligations. Extraterritorial political obligations might simply refer to political obligations that extend beyond the territory of the state. On Anna Stilz's account of political obligations someone who has a political obligation to a state owes the polity 1) obedience to the law, 2) participation in political debates, 3) taxes, and 4) contribution to welfare distribution.[1] Someone with extraterritorial political obligations would owe the state all or some of these things even when outside its borders.

The thesis of this paper holds for extraterritorial political obligations broadly construed. Accordingly, I believe that there is no liberal justification for requiring citizens abroad to obey the laws of their country of origin, pay taxes, vote, or carry out any civic duties that are part of their regular political obligation. However, I would face an extremely big burden of proof to show

that all extraterritorial political obligations are invalid. For example, I would have to respond to arguments made in favor of citizenship taxation (as opposed to territorial taxation) that are unique to the issue of taxation. Hence, I will limit the scope of the extraterritorial political obligations this paper is concerned with. When this paper refers to extraterritorial political obligations, it specifically addresses extraterritorial political obligations that can lead to punishment by criminal law when violated.[2] This excludes most civic duties such as voting, participating in political discourse, keeping an eye on government corruption, etc.

I further illuminate what I mean by extraterritorial political obligations by juxtaposing it with what it is not. Not all claims of extraterritorial jurisdiction are motivated by the need to enforce a political obligation nationals have to a state. We can think of a case analogous to the prosecution of Kim the Korean (insofar as it involves a state exercising extraterritorial jurisdiction), but different in the sense that it is not carried out on the grounds of extraterritorial political obligation. Child prostitution is illegal in the United States. Suppose that Jane, a U.S. Citizen, were to pay children for sex in a country where child prostitution was legal or child prostitution laws went unenforced. Under federal law, the U.S. still wields the legal authority to punish her for soliciting child prostitution outside its borders.[3] It is questionable, however, that Jane would be punished

on the grounds that she violated a political obligation shared by U.S. residents and citizens. It is more likely that the law is meant to punish Jane for violating a universal moral obligation. This subtle difference becomes more apparent when we think about who should punish Jane. If we think that committing sex offense against a child constitutes a grave transgression of such moral obligation, we would rejoice when Jane gets punished for paying children for sex abroad. We may be indifferent about the particular entity that penalizes Jane. It is very possible that we believe that any state, or even private entities, should be authorized to punish whoever commits such despicable crimes than see them go free. This view is best embodied by the phrase, "No safe haven for perpetrators of crimes against humanity."

The distinction between the violation of political obligation and the violation of moral obligation explains why this paper will scrutinize the justifications for the prosecution of Kim but not of Jane. This is not to say that political obligations lack a moral component to them. The consensus among political theorists is that "to have a political obligation is to have a moral duty to obey the laws of one's country or state."[4] However, Jane's moral obligation to not pay children for sex is far from a political one because it is owed to all persons simply qua persons, not to a particular state. It exists independently of the state or its laws.[5] She

would be obligated to uphold this duty even if there weren't any U.S. laws against child prostitution. It just happened to be the United States, a state of which she is a citizen, that claimed the jurisdiction to govern her behavior because of mainly practical, logistical reasons. The U.S. was in the best position to make sure that Jane is punished for her crime against humanity and prosecuted her accordingly.

Finally, political obligations should not be conflated with legal obligations. That based on the Korean constitution, Kim was under the legal obligation to refrain from smoking weed wherever he is in this world is a statement of social fact. But the fact that a person has a legal obligation to do X provides him with a moral reason to do X only in the case that he has a moral duty to obey the law.[6] Meanwhile, that a person has a political obligation to do X implies that this condition is satisfied--he had a moral duty to obey the law. This distinction between legal and political obligations makes it possible that a person is subject to a legal obligation even though she has no political obligation to obey the laws of the sovereign.[7] Most theorists will agree that people in a tyrannical, unjust state still have legal obligations to obey the laws of the state even though they owe no political obligation to the state.[8] This paper is not interested in the question of whether Kim was legally obligated to abide by the Korean law that prohibits him from smoking weed abroad. Rather, it is interested in whether he had an extraterritorial political obligation that imposed a moral duty to not smoke weed abroad.

Justifications for Extraterritorial Political Obligation

The history of political theory is riddled with struggles to provide a satisfactory account of political obligation. It is difficult enough to demonstrate that some kind of political obligation to states exists. Defending a political obligation to a particular state is even harder.[10,11]Various scholars have attempted to provide liberal justifications for what is called the particularity challenge.[12] I specifically focus on three theories that attempt to justify political obligation to particular states and assess whether they can extend justifications to extraterritorial political obligations. I charitably review each account, assuming that they are somewhat successful in demonstrating that political obligation to particular states exists in general.

1. Consent

One popular justification for political obligation to particular states is that members of the polity have consented to be subject to such an obligation. Most critics don't dispute that political obligations can be grounded on consent, but they find it unlikely that most people subject to political obligations have given express or tacit consent, or at least "the kind of ac-

tual consent that can ground a general obligation," to begin with.[13] They do not think hypothetical consent is good enough to generate political obligation.[14] In response to this criticism, defenders of the consent theory of political obligation have proposed various ways in which consent can be construed more broadly.[15] For example, some have argued that taking part in quotidian activities such as using the public library, voting in elections, and taking advantage of social security benefits can signify consent.[16]

This paper is less interested in how successful the consent theory is than what it says about extraterritorial political obligations if successful. Hence, I will assume that members of a polity have consented to some kind of political obligation to the state. Suppose that Kim the Korean retained his Korean citizenship even given the opportunity the renounce it. Moreover, he has a history of actively participating in the institutions of the state. We may think that such gestures of consent naturally entails an agreement to abide by the rule of law. Kim agreed to obey the laws of Korea, and the Korean law includes a duty to uphold all types of political obligations. It seems reasonable to conclude then that he agreed to uphold both domestic as well as extraterritorial political obligations. There appears little reason to separate the two types of obligations. Notwithstanding the broad interpretation of consent, however, I believe that there is still great difficulty in

claiming that political obligations justified under the consent theory include extraterritorial political obligations. One's consent to undertake political obligations to a state in general cannot be extrapolated to mean a valid consent to undertake extraterritorial obligations. I argue so on the basis that only ongoing consent constitutes valid consent. Extraterritorial political obligations preclude ongoing consent to an even greater degree than political obligations limited to state borders, casting doubt on the validity of the putative act of consent expressed by members of the state.

Putative acts of consent do not always generate obligations. Most theorists would believe that if the sovereign's authority or the nature of the obligation were unjust to begin with, the act of consent is void.[17] For example, a consent to be a sovereign's slave generates no obligation even if it genuinely expresses the subject's will.[18] Another instance in which an act of consent generates no obligation is when a person consents to an non-negotiable, permanent obligation. Even if the person wills to undertake such an obligation and never objects to its terms throughout the years, I believe it lacks liberal justification. Citizens are unable to exercise self-determination if they cannot assess the set of laws and obligations as a whole and decide to get out of it. The threat to self-determination posed by immutable terms of agreement is exacerbated by the fact that a state's laws and the precise con-

tent of the political obligation are subject to change all the time. It is true that democratic states allow citizens to take part in the formation of their laws; however, an individual has no choice but to put their fate on the hands of the majority. I think even defenders of consent theory would have to accept that putative acts of consent to irreversible terms comes short of being a valid consent.

An obvious objection to the argument so far is that most states allow members to renounce their citizenship at any point, and that given this opportunity to opt-out, retaining one's citizenship might suggest ongoing consent. However, the option to leave is really no option at all since it comes with the cost of becoming a stateless person who is stripped of any state-provided protection. It cannot be said that Kim is given a real choice to revoke his decision at any point beyond that. The choice is binding for life. At the very least, the purported ability to renounce one's citizenship at any point in time is not enough to guarantee ongoing consent.

One way in which states can make up for the difficulty of receiving ongoing consent from their members is by granting citizens the freedom of movement in and out of the state. The fact that a citizen remains within the state territory (when the option to travel abroad is readily available) can serve as additional affirmation of

their consent to obey the laws of the state. An additional layer of consent is established only if leaving the country means actual hiatus from being subject to the regular political obligations one owes the state. Extraterritorial obligations precludes this possibility and further undercuts the means of receiving ongoing consent from citizens because they follow the person regardless of where they are. It is problematic that when people are subject to exterritorial political obligations, a one-time putative act of consent (if there ever was one) generates permanent obligations that can neither be revoked or temporarily suspended.

2. Fair Play

Another prominent account of political obligation is what Rawls once referred to as the "duty of fair play."[19] The principle of fair play holds that everyone who participates in a reasonably just, mutually beneficial practice have an obligation to cooperate.[20] Free riders are considered to be doing wrong to the other participants of the shared enterprise even if its survival does not depend on their shirking. According to the fair play principle, they are obligated to share a fair burden in the enterprise because cooperation is what makes it possible for any individual to enjoy the benefits of the practice.[21] Hence, those who are part of the joint enterprise have rights against as well as obligations to one another: "a right to require others to bear their share of the burdens and an obligation to bear one's

share in turn."[22] The obedience to the laws of the state is deemed seminal to, perhaps even constitutive of, the maintenance of this cooperative enterprise. The members of the polity are thus obligated to uphold the rule of law under the fair play principle. In this manner, the principle of fair play provides grounds for a general obligation to obey the law.

Critics and advocates of the fair play theory alike have stipulated a few conditions that must be met in order for it to adequately provide grounds for political obligations. While some of the conditions are contentious, most political theorists agree that the principle of fair play applies to a political society only if its members can reasonably regard it as a cooperative enterprise.[23] It would be unreasonable to declare that a person has an obligation to take part in project a select group of people arbitrarily decided to take upon themselves. Second, the benefits of the cooperative practice must be of relatively substantive value to the participants.[24] If the products of the enterprise are of negligible value, then it is harder to defend an obligation (on the part of the members) to share the burdens of the enterprise. Third, even if the enterprise produces benefits, merely receiving benefits may not require someone to partake in the enterprise. Scholars who support this condition contend that receiving and accepting benefits are two different things.[25] To generate obligations, members must be aware that "the benefits are provid-

ed by a cooperative scheme" and that they could forgo those benefits in exchange of bearing no obligation to the enterprise.[26] In other words, there must be some indication that members consider those benefits worth undertaking a burden for. Fourth, the obligation of fair play applies to members of a polity only when their failure to obey the rules could affect the enterprise. This last condition is perhaps the most contentious out of all four.[27] Critics of the fair play theory have pointed out that "the obligation of fair play governs a man's actions only when some benefit or harm turns on whether he obeys."[28] Accordingly, they argue that the fair play theory cannot be applied to most polities because modern day states are simply too big to be affected by an individual member's actions. Advocates of the fair play theory have in response claimed that "fairness is not a consideration only when harm or benefit to some person or practice is involved."[29] They believe that to fail to do one's part in a cooperative enterprise is unfair and thus wrong to those who cooperate regardless of how this failure impacts the enterprise.[30]

It is difficult to assess whether the fair play theory can cover extraterritorial obligations among the political obligations it justifies. As with other theories, the theory can potentially show that members have political obligations to obey the law in general, but it says little about what specific laws can be included or excluded

on the grounds of fair play since they are altogether lumped into the rule of law. Perhaps whether fair play theory can justify extraterritorial obligation comes down to whether an enterprise that subjects members to extraterritorial political obligations meets a set of empirical conditions. I specifically tailor the four conditions described above such that they delineate the conditions on which fair play theory can ground extraterritorial political obligations.

1. Do the subjects of the extraterritorial political obligation regard themselves as being part of a cooperative enterprise?

2. Does the enterprise generate something valuable for its participants by requiring compliance to extraterritorial political obligations?

3. Do people consider upholding extraterritorial political obligations as part of their fair share in the cooperative enterprise? In other words, do people accept the tradeoff of being subject to extraterritorial obligations in return for what they believe are the benefits associated with them?

4. Does a person's disobedience to extraterritorial political obligations affect the common enterprise?

Some of the defenders of the fair play theory may protest that some of these conditions shouldn't even be considered.[31] It must be admitted that these conditions have been put forth by mostly critics of the theory.[32] However, I operate under the presumption that

some polities do meet these conditions and thereby are able to justify political obligations in general. I believe this sets up a favorable starting point for the case for extraterritorial political obligations.

The first and third conditions do not appear to pose significant problems for the defenders of extraterritorial political obligations on the grounds of the fair play theory. The first condition is the easiest to satisfy. Surely we can think of a polity (that happens to impose extraterritorial obligations on its members) whose members consider themselves involved in a joint enterprise. Korea, for example, would most likely pass this test. In addition, it is quite possible, if not likely, that most Koreans are content with the state imposing extraterritorial obligations on them. They accept extraterritorial obligations that say, bar them from smoking weed because they believe it contributes to keeping Korea drug-free, which they value. So the third condition is satisfied as well.

There is greater difficulty in showing that polities satisfy the second and fourth conditions, however. That individuals feel that being collectively subjected to extraterritorial obligations brings benefits, does not necessarily mean that it really does. But the objective fact of whether benefits are produced from the joint enterprise matters. Imagine a town that levies additional money from residents to support a water filtra-

tion system. It claims that the water filtration system is responsible for the town's clean tap water, something the residents value. Most of the residents happily contribute, convinced that the clean water filtration system helps maintain water quality. But unbeknownst to most villagers, the reality is that the city's factory waste laws are strict enough to keep tap water clean without the need for a water filtration system. It is hard to see how there is a real obligation to pay taxes when the project produces no benefits for its members. Moreover, the consideration of fairness (which, granted, has the ability to make the fourth condition void), cannot salvage the obligation in this case (i.e. when it comes short of meeting the second condition). Even if other villagers were to protest an individual's noncompliance to pay the additional tax on the basis of fairness, it seems wrong to demand an obligation from the individual when the obligation generates no benefits.

What this means is that justifications for extraterritorial political obligations must demonstrate the presence of real benefits, not simply the majority's perception of benefits. The link between extraterritorial political obligations and their claimed benefits are tenuous at best, however, given the extraterritorial nature of these obligations. They cannot be enforced properly because the state lacks the means to oversee their nationals while abroad. Hence this casts doubt on

their effectiveness as deterrence. In the case of Korea, for example, Korea would have to show that prohibiting Koreans from smoking weed abroad serves its goal of maintaining a drug-free state if that hasn't been achieved already by its draconian domestic drug laws.

Finally, it is also unlikely that the fourth condition is satisfied. The fourth condition is closely connected to the second, but places additional burden on the state looking to justify extraterritorial political obligations. The cooperative enterprise must not only generate real benefits as required by the second condition, but also be affected by the disobedience of an individual. We can imagine a case in which the second condition is satisfied but not the fourth. Even if Korea can successfully argue that prohibiting Koreans from using drugs abroad generally improves Korea's chances of remaining drug-free, it may be unable to show that a single individual's noncompliance endangers the cooperative enterprise. For one, it will have to adopt the premise that the individual will return to their homeland. Additionally, it will have to demonstrate that a person's disobedience triggers a chain of events substantial enough to leave an impact on the enterprise.

Perhaps overcoming these burdens of proof is not impossible. I do not completely rule out the fair play justification of extraterritorial political obligations as a whole. Not to mention, it is possible that extraterri-

torial taxation carries a greater potential to pass the same test than the extraterritorial duty to not smoke weed depending on the extent to its impact on the cooperative enterprise.[33] Nonetheless, the success of justifying extraterritorial obligations will be heavily contingent on a myriad of empirical assumptions, many of which are uncorroborated as of now.

3. Equal Freedom

This paper finally examines a liberal justification of special obligations to states put forth by Anna Stilz. Inspired by Kant's and Rousseau's theories, Stilz makes the case that the value of equal freedom can only be realized through the mediation of the state.[34] Absent the state, "the value of equal freedom is indeterminate" because there is no authority to set up a common legal order that people can reference.[35] And unless there is a public definition everyone can reference, people cannot be guaranteed a private sphere of liberty.[36] This is because when agents are left to act on their private judgments of justice, it becomes possible to unilaterally impose their private conceptions of justice on others.

As such, Stilz believes that an obligation to a particular state can be derived from natural duties. Equal freedom requires us to accept political authorities that define the rule of law. Is there an additional reason to owe obligations to a particular state? Stilz believes

that the simple fact that fellow compatriots engage in a collective cooperation can justify special obligations to them.[37] She believes that it doesn't require reference to shared history, culture, or language to ground special obligations, an approach that runs the risk of contradicting liberal ideas.[38]

I find Stilz's account of political obligations the most promising out of the three. But it doesn't appear to give sufficient grounds for extraterritorial political obligations. Stilz heavily relies on the presence of shared institutions among people to explain political obligations. Only through shared institutions can the value of equal freedom be realized. While the condition of shared institutions is not the same thing as a shared territory, they pretty much go together: "the duty to obey the law, if there is one, is an obligation that binds only those persons who stand in some special institutional relationship--those who fall within the territorial domain of a given state."[39] Hence, Stilz argues that a political duty to obey the law extends to foreigners who reside and work in the state. Even if they carry a foreign passport, they also ought to positively contribute to the state of residence, for it protects their equal freedom by issuing a common rule of law for all those within the territory. Conversely, states also have the duty to realize equal freedom for foreigners in their territory: "States themselves are justified institutions only because, and insofar as, they are necessary to realize

the equal freedom of individuals, and that means all individuals, not just their own members."[40]

In fact, because the proximity to institutions is a morally significant consideration for Stiltz, it is hard to locate in her account how states can justify imposing political obligations on nationals abroad. If indeed "equal freedom can only be defined and guaranteed within the [territories of the] state," it is doubtful that states at all contribute to the equal freedom of its nationals abroad.[41] It follows then that they cannot demand political obligations from national abroad on the basis of guaranteeing their equal freedom. Yet she seems to accept that foreigners residing abroad still have political obligations to their country of origin aside from obeying its laws (i.e. voting and fulfilling other civic duties). It is expected that these obligations are fulfilled remotely. It is questionable how she accounts for such political obligations that follow people no matter where they go.

One answer Stilz can give is that citizens and foreign residents have different levels of shared intentions with the members of the state (in which they both reside). Hence, requiring citizens to perform more demanding obligations is not a form of arbitrary discrimination. This response runs into the problem of relying on an individual's subjective sense of their relation to a collective. It cannot account for cases in which a

foreign resident still considers herself deeply involved in the collective cooperation with those around her. Additionally, it may be argued that a citizen abroad still receives benefits from her country of origin even if the benefits are not enough to guarantee equal freedom. For example, the U.S. government will try to extricate its citizens if they were held hostage abroad. However, this argument cannot justify taxing nonresident Americans similarly to resident Americans who receive far more benefits.[42]

The more convincing answer seems to be that Stilz does not intend to establish further moral differences between citizens and foreign residents nor justify extraterritorial political obligations. She doesn't say so explicitly, but it is quite likely that her theory supports a territory-based taxation over a citizenship-based taxation unless there are some prudential reasons for enforcing the latter. In short, there is a caveat to her case for liberal loyalty. It only applies when a person's state of origin matches the state of her actual residence. Granted, the fate of extraterritorial tax under Stilz's theory may be less certain, but it surely fails to defend extraterritorial obligation to obey the laws of the country of origin.

Conclusion

Simply because the three theories above cannot account for extraterritorial political obligations does

not mean they can ever be justified. For one, there may be theories this paper did not cover that can do the job. But also, there are plentiful illiberal grounds for extraterritorial political obligations. For example, nationalists argue that a reference to a shared national culture--common language, practices, myths, and territory, etc.--is integral to providing citizens with a special reason to support a particular state.[43] But contrary to the assertions of the self-proclaimed liberal-nationalists, the nationalist argument is not truly liberal.[44] This is because the argument only requires a self-claimed association with a group to generate special obligations to that group. This minimal requirement leads to problematic implications such as justifying special obligations to the Mafia simply because one is born into it or identifies with the group.[45] It is worrisome then that so-called liberal democracies borrow heavily from the nationalist justification of special political obligations to defend extraterritorial political obligations in the real world.

What the paper does show is that prominent liberal theories of political obligation cannot give good liberal reasons for extending political obligations to citizens abroad unless they employ flimsy empirical assumptions. For the time being, therefore, liberal states should find imposing extraterritorial political obligations inconsistent with their values. Moreover, the paper highlights a vacuum in political obligation

literature. Each account of political obligation is unclear about where they stand on this question of how much political obligation depends on territory. A common view seems to be that a successful defense of political obligation, even that to a particular state, does not hinge on the answer to that question. That may be true. The priority of political theorists has been on establishing some kind of political obligation in the first place. Thinking about its exact bounds is considered secondary to this more crucial task. Nonetheless, it is useful to think about what a particular account of political obligation, if successful, really justifies. This is rarely illuminated by the accounts and deserves further exploration.

Notes

1. Anna Stilz, *Liberal Loyalty: Freedom, Obligation, and the State* (Princeton: Princeton University Press, 2011), 209.
2. The evasion of extraterritorial taxation or citizenship taxation can lead to penal action. Hence, I don't exclude taxation from the scope of this paper, but there may be unique arguments in favor of citizenship taxation that cannot be extended to support other extraterritorial obligations. This possibility is further discussed in this paper's section on fair play theory.
3. Under 18 U.S.C. § 2423 (c) the U.S. claims "extraterritorial jurisdiction" over certain sex offenses against children, prohibiting not only citizens but also permanent residents from raping, sexually molesting, or soliciting sex from children while abroad.
4. Richard Dagger and David Lefkowitz, "Political Obligation" The Stanford Encyclopedia of Philosophy. Published Fall, 2014, accessed Dec 15, 2018, https://plato.stanford.edu/entries/politi-

cal-obligation/#FaiPla.
5. In the case of Kim, by contrast, we can be more confident that Korea's exercise of extraterritorial jurisdiction is grounded on the putative violation of a political obligation Kim owed to Korea, specifically. Korea, no matter how poorly it thinks of those who smoke weed, does not find it a universally recognized moral wrong. For a stronger support for this empirical claim, consider an analogous ban on gambling for Koreans that apply both within and outside its borders. Koreans are forbidden from gambling in Korea and abroad in all but one Casino in rural part of Korea. Interestingly, every single other Casino in Korea allows, or even encourages, gambling for foreigners. From this we can infer that clearly, Korea does not bar Koreans from gambling on the grounds that it is a universal moral wrong, but rather because it sees it as part of a political obligation all Koreans owe. In this sense, the distinction I draw between the case of Kim and that of Jane does not necessarily stem from the degree of immorality or criminality of their actions. It is different in a more fundamental way.
6. Ibid.
7. Ibid.
8. Ibid.
9. Ibid.
10, Michael Huemer, *The Problem of Political Authority: An Examination of the Right to Coerce and the Duty to Obey* (Basingstoke: Palgrave Macmillan, 2013).
11. Anna Stilz, *Liberal Loyalty: Freedom, Obligation, and the State*, 6.
12. Ibid.
13. Richard Dagger and David Lefkowitz, "Political Obligation."
14. A. John Simmons, "Liberal Impartiality and Political Legitimacy," Philosophical Books 34, No. 4, (1993): 220-21.
15. Peter Steinberger, *The Idea of the State* (Cambridge: Cambridge University Press, 2004), 218.
16. Richard Dagger and David Lefkowitz, "Political Obligation."

17. David Estlund, *Democratic Authority: A Philosophical Framework* (Princeton, NJ: Princeton University Press., 2008), 119-137.
18. Richard Dagger and David Lefkowitz, "Political Obligation."
19. John Rawls, "Legal Obligation and the Duty of Fair Play," in *Law and Philosophy*, ed. S. Hook (New York: New York University Press, 1964).
20. Richard Dagger and David Lefkowitz, "Political Obligation."
21. Ibid.
22. Ibid.
23. Ibid.
24. George Klosko, *The Principle of Fairness and Political Obligation* (Lanham, MD: Rowman & Littlefield, 2004) 38–9.
25. A. John Simmons, *Moral Principles and Political Obligations* (Princeton, NJ: Princeton University Press, 1979), 129.
26. Ibid., 132.
27. William Edmundson, *The Duty to Obey the Law: Selected Philosophical Readings* (Lanham, Md: Rowman & Littlefield, 1999).
28. Ibid., 81.
29. Richard Dagger and David Lefkowitz, "Political Obligation."
30. Ibid.
31. Richard Dagger, *Civic Virtues: Rights, Citizenship, and Republican Liberalism* (New York: Oxford University Press, 1997), 69-78.
32. See Nozick (1974) and Simmons (1979).
33. Some scholars like Mason (2016) have argued in favor of citizenship taxation on the grounds that "the failure to tax overseas Americans' total income, no matter where earned, would result in their systematic underpayment of taxes compared to resident Americans" (4). On this view, citizenship taxation ensures the equal treatment of resident and nonresident citizens.
34. Anna Stilz, *Liberal Loyalty: Freedom, Obligation, and the State*.
35. Ibid., 21.
36. Ibid.
37. Ibid., 8-9.
38. Ibid., 15-20.

39. Ibid., 4.
40. Ibid., 108.
41. Ibid., 96.
42. Ruth Mason, "Citizenship Taxation" *Southern California Law Review* 89, (2016): 4.
43. Anna Stilz, *Liberal Loyalty: Freedom, Obligation, and the State*, 139
44. For a longer critique of the liberal-nationalist account of special obligations, see Stilz 140-8.
45. Ibid., 146.

References

Blakesley, Christopher L. "United States Jurisdiction over Extraterritorial Crime," *Journal of Criminal Law & Criminology* 73 no. 3 (1982): 1109-1163.

Dagger, Richard. *Civic Virtues: Rights, Citizenship, and Republican Liberalism*. New York: Oxford University Press, 1997.

Dagger, Richard and David Lefkowitz. "Political Obligation," in The Stanford Encyclopedia of Philosophy, ed. Edward N. Zalta, 2014. Accessed Dec. 15, 2018. https://plato.stanford.edu/entries/political-obligation/#FaiPla

Edmundson, William. *The Duty to Obey the Law: Selected Philosophical Readings*. Lanham, Md: Rowman & Littlefield, 1999.

Estlund, David. *Democratic Authority: A Philosophical Framework*, Princeton, NJ: Princeton University Press, 2008.

Huemer, Michael. *The Problem of Political Authority: An Examination of the Right to Coerce and the Duty to Obey*, Basingstoke: Palgrave Macmillan, 2013.

Mason, Ruth. "Citizenship Taxation," *Southern California Law Review*, Vol. 89, 2016.

Nozick, Robert. *Anarchy, State and Utopia*. Basic Books, 1974.

Rawls, John. "Legal Obligation and the Duty of Fair Play," in *Law and Philosophy*, ed. S. Hook, New York: New York University Press, 1964.

Simmons, A. John. *Moral Principles and Political Obligations*. Princeton, NJ: Princeton University Press, 1979.

Simmons, A. John. "Liberal Impartiality and Political Legitimacy," *Philosophical Books* 34, no. 4 (1993): 213-223.

Steinberger, Peter. *The Idea of the State*, Cambridge: Cambridge University Press, 2004.

Stilz, Anna. *Liberal Loyalty: Freedom, Obligation, and the State*, Princeton: Princeton University Press, 2011.