Northwestern University
Society for the Theory of Ethics and Politics

Sixth Annual Conference
May 17-19, 2012

Conference Program

Program Photos: www.northwestern.edu/magazine/photogallery/index.html

Schedule

Thursday, May 17th, 2012

Location: John Evans Alumni Center

Morning Session

9:00-10:25. "Moral Realism, Evolution, and Our Reasons to Survive"
Speaker: Jeff Behrends, University of Wisconsin-Madison
Commentator: Guy Elgat, Northwestern University

10:35-12:00. "Moral Realism and Proper Function"
Speaker: Jeffrey Wisdom, Joliet Junior College
Commentator: Anne Eaton, University of Illinois, Chicago

Lunch

Afternoon Session

2:00-3:25. "The Problem with (Quasi-Realist) Expressivism"
Speaker: Steve Davey, University of Texas-Austin
Commentator: Raff Donelson, Northwestern University

3:35-5:00. "Coercion and Moral Explanation"
Speaker: Arudra Burra, UCLA
Commentator: Japa Pallikkathayil, NYU

Dinner at 6:00pm

Friday, May 18th, 2012

Location: Harris Hall, Rm. 108

Morning Session

9:00-10:25. "Free Will, Consequential Responsibility and the Concept of Distributive Justice"
Speaker: Attila Mraz, NYU and CEU
Commentator: Tyler Zimmer, Northwestern University

10:35-12:00. "Action as Interaction"
Speaker: Kristina Gehrman, Miami University of Ohio
Commentator: Scott Forschler, St. Cloud Technical and Community College

Lunch

Friday, May 18th, 2012 (cont.)

Location: Harris Hall, Rm. 108

Afternoon Session

2:00-3:25. "Empathy, Proper Empathy, and Understanding"
Speaker: Yujia Song, University of North Carolina-Chapel Hill
Commentator: Lee Goldsmith, Northwestern University

3:35-5:30. Keynote Address: "Volitional Rationality and the Necessities of Love"
Speaker: Harry G. Frankfurt, Princeton University
Commentator: Andrea Westlund, University of Wisconsin-Milwaukee

Reception – Everyone is invited – 6:30pm

Saturday, May 19th 2012

Location: Harris Hall, Rm. 108

Morning Session

10:35-12:00. "Rational Requirements and 'Rational' Akrasia"
Speaker: Edward Hinchman, University of Wisconsin-Milwaukee
Commentator: Jesse Summers, Duke University

Lunch

Afternoon Session

2:00-3:25. "Promises, Practices, and Interpersonal Obligation"
Speaker: Jorah Dannenberg, Stanford University
Commentator: Erin Taylor, Cornell University

3:35-5:30. Keynote Address: "Ideas of the Good in Moral and Political Philosophy"
Speaker: T. M. Scanlon, Harvard University
Commentator: David Sussman, University of Illinois at Urbana-Champaign

Dinner at 7:00pm

Table of Contents

Speaker Biographies and Papers (pp. 5-104)

Speaker Biographies and Papers

(in order of appearance on the program)

Jeff Behrends

University of Wisconsin-Madison

Jeff Behrends is a doctoral candidate at the University of Wisconsin-Madison. His recent publications include articles on the nature of relational goods, and on the role that the notion of *promoting a desire* plays in practical rationality. He is currently working on a dissertation regarding the relationship between desires, value, and practical reasons, in which he defends the position that practical reasons are generated both by the desires of agents and by certain kinds of value.

"Moral Realism, Evolution, and Our Reasons to Survive"

Abstract: *In this paper, I attempt to explain and respond to the evolutionary challenge for moral realism. The challenge, roughly, is that of explaining how our moral judgments could come to be adequately correlated with the moral facts, given certain evolutionary facts, and assuming the truth of moral realism. If the challenge cannot be met, then moral realists are committed to moral skepticism, and for that reason realism should be rejected in favor of some other meta-ethical position. I examine David Enoch's (2010, 2011) proposed solution to this challenge, and argue that it is in adequate. I then argue that Enoch's strategy can be successfully implemented in a new way. In short, I argue that, if moral realism is true, it would be a kind of fantastic coincidence if a significant amount of our evolutionarily-influenced moral judgments were not correctly correlated with the moral facts.*

1. Introduction

It is commonly thought that moral realism faces a variety of epistemological challenges, but few have received as much attention in recent years as the evolutionary arguments offered by Sharon Street (2006) and Richard Joyce (2006).[1] Any meta-ethical position that recognizes the existence of moral facts must involve some account of how our moral judgments can, at least sometimes, come to be correctly correlated with those facts. If evolutionary forces are likely to have significantly influenced our moral and evaluative judgments, then what evidence could we have for thinking that those judgments are indeed correlated with the truth? It is easy to see how various constructivist positions can accomplish this, but the evolutionary challenge raises legitimate concerns on this front for the realist.

In this paper, I attempt to defend moral realism against the evolutionary challenge.[2] In the first section, I explain in more detail exactly what I take the challenge to be. In section 2, I consider and

[1] For a sample of discussions that those works have given rise to, see Copp (2010); Enoch (2010, 2011); Skarsaune (2011); Street (2008); and Wielenberg (2010).

[2] As Enoch (2010) correctly observes, the evolutionary challenge is significant not only for meta-*ethical* realism, but also for meta-*normative* realism more generally. As the challenge has usually been discussed as a meta-ethical argument, though, I will discuss it as such here. I take it, though, that almost everything I say about the evolutionary challenge and its

[2] As Enoch (2010) correctly observes, the evolutionary challenge is significant not only for meta-*ethical* realism, but also for meta-*normative* realism more generally. As the challenge has usually been discussed as a meta-ethical argument, though, I will discuss it as such here. I take it, though, that almost everything I say about the evolutionary challenge and its

criticize David Enoch's recently proposed solution to the challenge. I argue that although Enoch has identified a promising strategy for realists to take, he has failed to adequately implement it. In the remaining sections I offer what I take to be a superior implementation of the strategy, followed by some concluding remarks on the state of the debate if my argument is successful.

2. What is the Evolutionary Challenge to Moral Realism?

I agree with Enoch (2010, 2011) that the evolutionary challenge to moral realism is really a particular version of a more general epistemological challenge to the position. In short, the challenge that the realist must meet is that of explaining how our moral judgments could come to be correctly correlated with the moral facts, realistically construed.[3] The challenge becomes more imposing once we attend to the genealogy of our moral judgments, for it appears that we have come to hold those judgments for reasons that are completely independent of the content of the moral facts. Let us call this challenge to moral realism the *Correlation Problem*.

The evolutionary challenge begins with the identification of one important causal force in the genealogy of our moral and evaluative judgments: evolutionary selection pressure. Other versions of the Correlation Problem might focus on other genealogical influences: social customs and practices, religious indoctrination, or the moral attitudes of our parents, for example. The evolutionary version of the Correlation Problem starts with the following plausible hypothesis: evolutionary forces have had a significant impact on the moral and evaluative judgments we endorse. Call that hypothesis the *Evolutionary Claim*. Note that, if the Evolutionary Claim is true, it will be a matter of some interest exactly *why* it is true. How have evolutionary forces impacted the content of my moral judgments? Have particular judgments been subject to selection pressure? Have judgment-forming systems been selected for instead? While the answers to these and similar questions are no doubt important in some respect, I want to leave them unexamined for the purposes of this discussion. It is enough to get the evolutionary challenge going that the Evolutionary Claim is true, and I am willing to grant that it is.

Given the truth of the Evolutionary Claim, the realist appears to be in the position of needing to supply an explanation for the relationship between the forces of evolution and the stance-independent moral facts. As Street (2006) sees it, the realist is faced with a dilemma here. On the first horn of the dilemma, the realist could say that there just is no important relation between these two things. The moral facts are what they are, and the forces of natural selection have nothing at all to do with them. The difficulty with this view, Street contends, is that it renders the Correlation Problem devastating for moral realism. If the selective forces that have significantly influenced the content of our moral judgments have nothing at all to do with the stance-independent moral truths, then we have little reason for thinking that the two should be correlated. In fact, it would seem to be fantastically *coincidental* if any strong correlation existed.[4] It would be as if we had simply taken a moral-belief-pill, which has caused us to have some moral judgments or other. In such a scenario, any justification we might have thought we possessed for our evaluative judgments would be undercut by our knowledge that we have taken the pill, for we would lack reason for thinking that the belief-forming mechanism is one that is reliable. Similarly, on this horn of the dilemma, whatever justification we might think we possess for believing our moral judgments to be stance-independently true is undercut by our knowledge that the content of our judgments has been largely determined by natural selection.

---

relationship to meta-ethical realism applies *mutatis mutandis* to the relationship between the evolutionary challenge and meta-normative realism. See, though, fn. 8 for a discussion about a move in the argument that must be handled differently by meta-ethical realists than it can be by meta-normative realists.

[3] What is it for a moral fact to be realistically construed? Following Russ Shafer-Landau (2003), I will say that for a moral fact to be true in the way a realist construes it is for it to be *stance-independently* true. That is to say that they are true, and that they are so not in virtue of their endorsement from any actual or hypothetical stance, position, or perspective.

[4] For a related line of reasoning involving moral realism and the possibility of a coincidence of this kind, see Bedke (2009).

On the second horn of Street's dilemma, the realist could argue that the mechanisms that result in our evaluative and normative judgments have been selected for by evolutionary forces *because they are truth-tracking*, in something like the same way that our perceptual systems have been selected for because they produce judgments that accurately reflect physical truths. If that were the case, then the Correlation Problem could be neatly disposed of; our moral judgments and the moral facts would be closely correlated because we would have been pushed by selective forces to endorse moral claims that are actually true, just as we have been pushed by selective forces to endorse perceptual claims that are actually true. The difficulty here, Street argues, is that this response is scientifically implausible. Whereas it is easy to see that judgments about the physical world will usually be fitness-enhancing to the extent that they are true, it is not at all obvious why the fitness-enhancing moral and evaluative attitudes must also be *true* in order to be fitness-enhancing. For example, the judgment *that killing my children is wrong* will be fitness-enhancing whether or not it is true, whereas the judgments about physical objects in our environment will be fitness enhancing only to the extent that they get things right.

On either horn of the dilemma, then, things do not look good for the moral realist, as the evolutionary challenge appears to constitute a very serious undercutting defeater for any moral claim understood realistically. It does so by casting doubt on the ability of our moral belief-forming mechanisms to reliably cause judgments that correlate with the moral facts, if we are to understand the moral facts as stance-independently true.

3. Denying the Dilemma: Enoch's Third Factor Explanation

Enoch (2010, 2011) has recently defended moral realism against the evolutionary version of the Correlation Problem. He begins by arguing that Street's dilemma for the moral realist is a false one. As Street sees it, once the realist realizes that opting for the first horn of the dilemma renders a solution to the Correlation Problem impossible, she must attempt to avoid that problem by claiming that our moral capacities have been selected for *because they are truth-tracking*. It is a mistake, though, to think that this is the only way of explaining the relationship between evolutionary forces and the moral facts while also addressing the Correlation Problem. It may be that selective forces have favored certain evaluative attitudes and judgments precisely because they are fitness enhancing, but that some additional fact explains why those attitudes and judgments correspond to the truth.[5] A successful response of this kind would explain the correlation without appealing to miraculous coincidences or scientifically implausible claims about our evolutionary history.

Enoch's (2010, 2011) idea is that our moral judgments are reliably correlated with the moral facts because a core evaluative claim that we have come to endorse as a result of natural selection is in fact true. Consider the judgment that our survival is good, or at least good *for us*. This is just the sort of judgment that is likely to have been subjected to strong selection pressure. We can expect that, other things being equal, creatures that look favorably on their own survival, and take there to be reasons that count in favor of actions that promote their survival, will be more reproductively successful than creatures that lack such attitudes or possess contradictory attitudes.

Suppose that it is true that our survival is good for us, and that we have come to endorse that claim as a result of evolutionary forces. How would this help realists solve the Correlation Problem? Enoch's idea is that so long as some core set of our normative judgments are correctly correlated with the stance-independent moral truths, ordinary reasoning will render us capable of strengthening that correlation to the point that the Correlation Problem can be dismissed. His contention is strengthened by his further observation that the correlation that realists must be capable of explaining isn't *perfect correlation*. Insisting that realists account for perfect correlation between our moral judgments and the stance-independent truths would be appropriate only if realists were committed to the claim that we

[5] For authors who advert to a similar strategy see Skarsaune (2011), and Wielenberg (2010).

have perfect moral knowledge. But, of course, no realist is actually so committed. For the purposes of this paper, I want to buy this part of Enoch's argument more or less wholesale. At the very least, I want to assume that if the moral realist is capable of explaining why we are justified in believing that there is *some* correlation between the content of our moral judgments and the moral facts, then she will have significantly weakened the epistemological significance of the Correlation Problem.

While Enoch's strategy for resisting the evolutionary challenge is promising, his execution of that strategy is flawed. Recall that his argument depends on the evaluative claim that our own survival is good. Enoch is aware that one way to object to his argument is to argue that his assumed evaluative claim is false, rendering his reply ineffectual. Enoch anticipates this kind of reply, and borrows from Knut Olav Skarsaune (2011) in addressing it. Of his assumption that survival is good for us, he writes:

> either this (quite plausible) assumption is true, or it isn't. If it is, the suggested explanation of the correlation works. If it is not, then the suggested explanation fails, and then [moral realism] may be committed to skepticism. But if such a highly plausible premise is false, perhaps skepticism is precisely the way to go here – as Skarsaune puts the point, if we don't know, then we don't know. Either way, then, the suggested way of coping with the epistemological challenge gives the right result.[6]

Enoch's reasoning here is clear: if it's true that our survival is good for us, then the Correlation Problem can be solved; if it's not true, then we ought to be moral skeptics anyway.

That reasoning, however, is flawed. Suppose it's true that if our survival is good for us, and moral realism is true, then our moral judgments will be adequately correlated with the moral facts. Should this console anyone who is worried that, in light of the Evolutionary Claim, we lack reason to think that our moral capacities are reliable? Presumably not. For they will insist that Enoch has relied on a moral claim, and since we are not yet in a position to trust our moral capacities, we don't yet have any reason to think that the claim is true. Notice that the concern is not about whether the normative claim is *actually true or false*; it's about whether, given the dialectical situation, we're justified in taking it to be true. What Enoch has shown is that, as long as it's true that survival is good for us, and moral realism is true, then we'll have lots of true moral beliefs. But he has not given us reason to think that that is how things are. If you do not think that the latter bit of knowledge is important, think again about the situation we would be in if we were to have taken a moral belief pill. Suppose we grant that so long as the moral belief pill gets us to a true moral belief to begin with, we'll be able to reason our way to other true moral beliefs. Would our position then be epistemologically satisfying if we were to find out that we had taken the pill? Surely not, for we would have no reason to confident that any of our moral judgments *were* correct to begin with.[7]

4. Our Reasons to Survive

As we have seen, the difficulty with Enoch's proposed solution to the evolutionary challenge is that it relies on a claim that it is dialectically problematic for realists to simply take as true without argument. However, the general strategy of Enoch's solution is so attractive that it is worth thinking about whether there is a non-problematic way to carry it out. In the remainder of this paper, I will argue that there is such a way. I will proceed by offering an argument that concludes with a claim that is very similar to Enoch's claim about the goodness of survival, but which is entailed by premises all of which are dialectically permissible for a realist to assert when addressing the evolutionary challenge. Following that, I critically assess each of the premises, defending some from objections. While I ultimately conclude that the argument I present is unsound, the *way* in which it is unsound provides evidence that the evolutionary challenge, and indeed the Correlation Problem in general, are *much* weaker than anti-realists have taken them to be.

---

[6] Enoch (2011), p. 171 – 172.

[7] My criticisms of Enoch's position have been significantly influenced by several discussions on the topic with Justin Horn.

Here is the argument that I wish to examine. I call it the Argument for Survival:

1. If we have reason to do something or other, then we have reason to pursue the means necessary to doing it.
2. Our existence is necessary for our doing anything.
3. Therefore, if we have reason to do something or other, then we have reason to pursue our own existence.
4. If moral realism is true, then we have reason to do something or other.
5. Therefore, if moral realism is true, then we have reason to pursue our own existence.

If the argument is successful, then its conclusion can be used to play the same role that Enoch's original claim about the goodness of survival was intended to. What explains how the content of our moral judgments could come to be correlated with the moral truths? Well, evolutionary forces have pushed us toward holding the belief that we have reason to pursue our own survival, and as it turns out, the truth of moral realism entails that we do have such a reason. That is because we must exist in order to fulfill whatever substantive demands the moral facts place on us.

The argument is valid. The unsupported premises are 1, 2, and 4. As we shall see, the most interesting of these is premise 2, and for that that reason I turn to evaluation of it only after examining premises 1 and 4. Again, I intend to argue that even though the argument is ultimately unsound, an understanding of its problems uncovers a promising realist reply to the evolutionary challenge.

5. Premise 1

Suppose that Katie has a reason to go to Seattle on Friday, and that the only means available to her of accomplishing this is to take the train. Does Katie thereby have a reason to take the train? I believe that she does. Premise 1 in the Argument for Survival is simply a statement of the intuitive idea that when we have a reason to Φ, we thereby have a reason to do the things that are necessary for us to Φ. Notice that premise 1 does not entail that when we have a reason to Φ, we thereby have an *all-things-considered* reason to pursue the means necessary to Φ. That reason, like the reason to Φ itself, could be outweighed by competing considerations. Notice also that premise 1 is not a statement of the familiar idea that when an agent has a desire, she thereby has a reason to pursue the means necessary to satisfying that desire. Premise 1 is neutral with respect to the grounding of normative reasons; one need not be a Humean about reasons in order to endorse it.

It might be thought that, just as it is problematic for the realist to assert that our own survival is good when responding to the evolutionary challenge, so too is it impermissible for her to assert premise 1. It is tempting to think this, given that premise 1 is a claim that has to do with normative concepts, and given that the evolutionary challenge is meant to call into question the realist's justificatory support regarding normative claims. I think, though, that it is not dialectically inappropriate for the realist to help herself to premise 1.

Premise 1 is a claim that has to do with the nature of normative reasons, and is not itself a substantive normative claim. Our intuition that premise 1 is true is an intuition of the sort that Michael Huemer (2008) has called "formal intuitions." Huemer describes formal intuitions as those that "impose formal constraints on ethical theories, though they do not themselves positively or negatively evaluate anything." What is important about formal intuitions is that they are unlikely to be the results of potentially distorting genealogical influences, such as evolutionary forces. The truth of premise 1 seems to be required by the very nature of normative reasons, by the concept of something's being a reason. While it is plausible that selection pressures may have thrown us off track with respect to judgments about concrete evaluative and normative judgments, it does not seem plausible that those pressures may have distorted our understanding of *what it is* for something to be a reason, and the conceptual consequences that follow. If they have, then this is a problem for *everybody*, not simply realists. Since everyone should accept premise 1, then, it is not dialectically impermissible for the realist to assert it in this context.

6. Premise 4

Premise 4 in the Argument for Survival is the claim that if moral realism is true, then we have reason to do something or other. What is meant by this claim? As throughout the argument, the kinds of reasons in question here are normative reasons for action. Normative reasons for action can be understood in the familiar way – as facts (or properties of actions, or whatever) that count in favor of, or justify, performing certain actions. So premise 4 amounts to the claim that if moral realism is true, then there exists at least one fact that counts in favor of some action or other on our part.

Why think that such an entailment holds? We should think this because, according to moral realists, the moral facts just *are*, or at least entail, normative reasons. According to contemporary realists, like Russ Shafer-Landau (2003) and Enoch (2011), the moral facts are *intrinsically reason-giving*. It follows, then, that if there are any moral facts out there of the sort that realists have in mind, then we will have at least some normative reasons for action.[8]

It is important that premise 4 makes no mention of any determinate action that we have reason to do. If it did, it would run afoul of the difficulty that Enoch's claim about the goodness of survival did. The truth of premise 4 is compatible with any understanding at all of what the stance-independent moral truths might require of us. If morality is primarily a business of bringing about certain states of affairs, then we will have normative reason to act in ways that promote such states. If instead morality demands that we develop certain dispositions of character, then we will have normative reason to act in ways that promote such a cultivation. Similar comments apply to alternative understandings of what the substantive content of moral facts is. What matters for premise 4 is not the *kind* of actions that moral facts provide reasons for, just *that* there will be actions of some kind or another that the moral facts count in favor of.

7. Premise 2

I have said already that I take premise 2 to be the most interesting premise in the Argument for Survival. On first glance, though, it might appear to be totally innocuous. Of course our existence is necessary for our doing anything – if we weren't around, after all, we couldn't do much of anything! That, of course, is exactly right. But such a defense of premise 2 relies on a misunderstanding of what that premise must really be about if the Argument for Survival is on track.

To see that this is so, consider the following case:

Stop Existing Moral realism is true, but the only reason that any agent ever has is to kill herself.

In Stop Existing it is a stance-independent truth, let us imagine, that we all have most reason to end our lives; morality *demands* it. If such a situation is conceptually possible, then the conclusion of the Argument for Survival is false. In this odd scenario, moral realism would be true, but no one would have any reason to pursue their own survival; everyone would have overriding reason to pursue just the opposite, in fact. The Argument for Survival is valid, though, and we have already seen that premise 1 is just a part of our concept of having a normative reason, and that premise 4 follows from the best version moral realism. So something must be off with premise 2.

Think again about the reason in Stop Existing. While it's true that in order to have a normative reason to end one's existence one must exist, it isn't true that such a normative reason gives any agent a reason to continue existing. In order to successfully supplement Enoch's solution to the evolutionary

[8] I should note, though, that is kind of position is not uncontroversial among moral realists. I am here assuming that the best version of moral realism is one that endorses *moral rationalism*, though that position is distinct from realism itself. I avoid direct attention to moral rationalism here as a defense of it would take me too far afield. A second point is worth noting, though. If we were to run a parallel version of the Argument for Survival its terms of practical meta-normative realism, rather than meta-ethical realism, there would be no corresponding controversy concerning the parallel version of premise 4. That is because practical meta-normative is, in part, an existence thesis about practical reasons.

challenge, we should be aiming for the conclusion that we have reason to pursue our *persistence* – that we have reason to act such that we *continue to survive*. But read with that understanding of 'existence,' premise 2 appears to come out false. It isn't true that our persistence is necessary for our doing anything at all; my continued persistence isn't necessary for my causing myself not to exist. So, strictly speaking, we ought to say that premise 2 is false.

8. Implications

Premise 2 of the Argument for Survival is false. But what of it? Many of the reasons that we typically take ourselves to have do require our persistence. If you have reason to care for your children for some extended period of time, for example, then you'd better stick around if you have any hope of acting in accordance with that reason. The point generalizes to almost any other kind of reason we could think of. No matter what the specific content of the normative facts are, so long as they entail that I have reason to do something at some future time, then it will follow that I have a reason to pursue my persistence. But the ways in which the moral facts could be arranged such that I *don't* have any reason to do anything in the future seem remarkably few. Stop Existing is one such scenario, but it is not at all clear to me that many other scenarios of this kind are possible.

How could a realist argue for the point that it is extremely implausible that the moral facts could be arranged in such a way as to entail that we have no reason to pursue our own persistence? She might attempt to appeal to another formal intuition. Consider the following claim: There is a *pro tanto* reason to make things better than they are now. Let us call this principle *Do Better*. I am tempted to say that Do Better follows from an appropriate understanding of reasons and what it is for something to be better than something else. Notice that the claim involves no substantive commitments regarding what things are of value, no commitments about what it would be for things to be better than they are now. It is simply the claim that the fact that things would be better if I Φ counts in favor of my Φing. The way in which things could be better will depend on the actual content of the evaluative facts, whatever they are.

If Do Better is true, and moral realism is true, then it is almost certainly the case that we have stance-independent reason to pursue our own persistence. This is because Do Better will entail that we always have some reason to continue acting, unless we have reached the maximally good state of affairs. If that is so, then the realist can contend that the only condition under which moral realism is true and evolutionary forces have distorted our moral judgments is the condition under which the current state of affairs is maximally good. Enoch is right that the realist can do much to diffuse the evolutionary challenge if she can show that we have good reason to think that at least some of our evolutionarily caused normative judgments would correspond to the stance-independent moral truths. I have suggested that we *do* have such a reason. We should think that the correlation exists because the only condition under which it would not is the condition under which the current state of affairs is maximally good. But, to turn the evolutionary challenge on its head, it would be almost *miraculous* if that turned out to be the case. So, we have reason to think that evolution has distorted our moral judgments only if a sort of miracle has occurred.

Even if Do Better is false, it seems to me that we have good reason to think that we will nevertheless have reason to pursue our own persistence if meta-normative realism is true. All that is required is that *some* normative fact counts in favor of our performing actions that we cannot perform by dying. Do Better is merely an extremely plausible example of such a normative fact that realists can advert to if pressed to supply an example. Many other normative facts would fit the bill, though. So few would fail to fit the bill, in fact, that the truth of moral realism almost guarantees that we have a reason to pursue our own persistence.[9]

[9] Although I did not intend for it to, it has been pointed out to me that my response to the evolutionary challenge bears a structural resemblance to Mark Schroeder's (2007) response to the Too Few Reasons problem for Humean theories of

It might be objected at this point that even if I have demonstrated that realists are capable of defending the claim that we have reason to pursue our own persistence, I have not gained the position much epistemological ground. For all I have said, it could be that the stance-independent moral facts are quite bizarre, even if they happen to entail that I shouldn't seek out my own death. So, the worry goes, in attempting to solve the evolutionary challenge, I have merely called attention to a perhaps weaker epistemological problem for realist, but one that is nonetheless very serious: Even if we aren't forced to *skepticism* by moral realism, we may be forced to an objectionably impoverished amount of moral knowledge.

The objection is a serious one, but I think that it ultimately rests on a failure to apprehend the significance of undermining the Correlation Problem. Consider the dialectical state of affairs prior to the introduction of that problem. Realists disagree with one another about how we secure justification for our normative claims, but they all agree that they must provide *some* account of how that happens. Some might opt for a coherentist position, others for reliabilism, and still others might fall back on the notion of self-evident beliefs. The options are several, and they can be combined with one another to form stronger epistemological foundations for realism. Notice that the Correlation Problem is not meant to cause problems for any of these epistemic theories *as such*. That is, the Correlation Problem does not call into question coherentism, or reliabilsm, or the existence of self-evident propositions. Rather, the Correlation Problem is meant to threaten meta-normative realism with the prospect of an *undercutting defeater*. Undercutting defeaters are meant to eliminate justification *which is ordinarily good*. Suppose, for example, that you are a meta-normative realist who maintains that our justification for our normative beliefs is built up from normative intuitions. If the Correlation Problem were insurmountable, then we would have to treat normative intuitions as failing to provide excellent justification, even if we are otherwise convinced that intuitions are in general capable of doing so. If the potential undercutting defeater is shown to be no defeater at all, though, we can help ourselves to whatever initial justification we were thought to have in the first place.

My point is that, so long as the response I've given to the evolutionary challenge is sufficient to refute it, it's not all that interesting that moral realism is consistent with a whole range of normative facts. This is because, by refuting the potential undercutting defeater, and absent other kinds of objections, realists get to help themselves to the justificatory story that they take to be most plausibile. And it is *those* justificatory stories that will explain why we have reason to think that the normative facts are one way, rather than another. Of course, it is open to anti-realists to call into question coherentism, or reliabilism, or self-evidentialism, or whatever, but objections that are meant to do so will be wholly independent from the evolutionary challenge and the Correlation Problem. Once the Correlation Problem is dispensed with, the realist will simply recommend that we proceed with actual normative theorizing, whatever that looks like. And it is that project that will provide us with epistemic reason for thinking that the moral facts have a particular content, as opposed to any other.[10]

References

---

normativity. In an attempt to explain how Humeans can account for the existence of agent-neutral reasons, he writes that such reasons are "*massively* overdetermined. They are reasons for anyone, no matter what she desires, simply because they can be explained by any (or virtually any) possible desire." Similarly, I have attempted to argue that there is a sense in which our reason to pursue our own persistence is massively overdetermined, if moral realism is true; it can be explained by virtually any way that the substantive moral facts could be. Interestingly, I believe that Schroeder's overdetermination response to the Too Few Reasons problem fails, though for reasons that do not apply to my own overdetermination argument. For an excellent discussion of Schroeder's argument, see Shafer-Landau (2012) and Schroeder (2012).

[10] My thanks to the following for helpful feedback or conversation related to this project: Russ Shafer-Landau, Michael Titelbuam, Sarah Paul, John Bengson, Justin Horn, Gina Schouten, Michael Roche, Casey Hegelson, Matthew Kopec, Holly Kantin, James Sage, Erik Wielenberg, and audiences at both the University of Wisconsin-Oshkosh and Depauw University.

Bedke, Matthew S. (2009). "Intuitive Non-naturalism Meets Cosmic Coincidence." *Pacific Philosophical Quarterly* 90: 188 – 209.
Copp, David. (2008). "Darwinian Skepticism about Moral Realism." *Philosophical Issues*. 18: 186 – 206.
Enoch, David. (2010). "The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it." *Philosophical Studies* 148: 413 – 438
Enoch, David. (2011). *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press.
Joyce, Richard. (2006). *The Evolution of Morality*. Cambridge: The MIT Press.
Schroeder, Mark. (2007). *Slaves of the Passions*. Oxford: Oxford University Press.
Schroeder, Mark. (2012). "Reply to Shafer-Landau, McPherson, and Dancy," *Philosophical Studies*, Volume 157, Number 3, pages 463 – 74.
Shafer-Landau, Russ. (2003). *Moral Realism: A Defence*. Oxford: Oxford University Press.
Shafer-Landau, Russ. (2012). "Three problems for Schroeder's hypotheticalism," *Philosophical Studies*, Volume 157, Number 3, pages 435 – 43.
Skarsaune, Knut Olav. (2011) "Darwin and moral realism: survival of the iffiest." *Philosophical Studies* 152: 229 – 243.
Street, Sharon. (2006). "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127: 109 – 166.
Street, Sharon. (2008). "Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying About." *Philosophical Issues*. 18: 207 – 228.
Wielenberg, Erik. (2010). "On the Evolutionary Debunking of Morality." *Ethics* 120: 441 – 464.

Jeffrey Wisdom
Joliet Junior College

Jeff Wisdom (Ph.D., Connecticut) is an assistant professor of philosophy at Joliet Junior College in Joliet, Illinois. He specializes in metaethics, and he is particularly interested in connections between metaethics, the philosophy of mind, and the philosophy of religion. Some of his recent work has appeared in *Philosophical Studies*, *The Southern Journal of Philosophy*, and *Metaphysica*.

"Moral Realism and Proper Function"

Abstract: *A common line of thought in metaethics is that certain facts about the evolutionary history of humans make moral realism implausible. Two of the most developed evolutionary cases against realism are found in the works of Richard Joyce and Sharon Street. In what follows I argue that a form of moral realism that I call proper-function moral realism can meet Joyce and Street's challenges. I begin by sketching the basics of proper-function moral realism. I then briefly sketch what I take to be the essence of Street's and Joyce's objections and show how proper-function realism answers them.*

INTRODUCTION

A common line of thought in metaethics is that certain facts about the evolutionary history of humans make moral realism implausible. Two of the most developed evolutionary cases against realism are found in the works of Richard Joyce and Sharon Street. In what follows I argue that a form of moral realism that I call proper-function moral realism can meet Joyce and Street's challenges. I will begin by sketching the basics of proper-function moral realism. I will then briefly sketch what I take to be the essence of Street's and Joyce's objections and show how proper-function realism answers them.

THE MORAL SENSE

Humans have the ability to make moral judgments. This ability is virtually universal and it seems to have endured because it helps humans engage in reproductively advantageous behavior. Further, our ability to make moral judgments likely relies on the operation of several interconnected areas of the brain.[11] For ease of reference, call the neurally-implemented mechanism or set of mechanisms responsible for our ability to make moral judgments our *moral sense*. Typically, a mechanism whose operation conveys survival value on the organisms that have it is thought to have a biologically-proper function.[12] More specifically, the proper function of a mechanism or trait is whatever that mechanism or trait does that tends to contribute to the fitness of the organisms that possess the mechanism or trait. Thus, the proper function of the moral sense will therefore be some activity—or, more likely, a somewhat complex set of activities—that tend to contribute to the fitness of the organisms that possess it.

Ostensibly, the moral sense confers reproductive advantage on the organisms that possess it by doing at least three related things; namely, motivating evolutionarily advantageous behavior, producing moral judgments, and producing mental states with a certain phenomenological or emotional content.[13] On just about any metaethical view represented in the literature, moral judgments

[11] For a survey of some of the more plausible candidate neural areas involved in moral judgment, see Casebeer and Churchland (2003).

[12] See, e.g., Woodfield (1976), Griffiths (1993), Millikan (1989), and Neander (1991).

[13] Moral antirealists like Joyce, Blackburn, and Gibbard make similar claims. As Street notes, "Blackburn…writes that an evaluative attitude's 'function is to mediate the move from features of a situation to a reaction'" Blackburn (1993), p. 168;

motivate behavior at least contingently, if not necessarily. Further, when viewed from the first-person perspective it seems as though the moral sense motivates behavior by producing in us a sense of "to be doneness" when we consider certain actions, and a sense of "not-to-be-doneness" when we consider other actions. It also seems plausible that whether a particular moral judgment tends to benefit or hinder a person depends on the content of that judgment. For example, consider the judgment, *it is forbidden to share food with one's kin*. In the vast majority of circumstances that humans have faced throughout history, operating by this judgment will be less advantageous than the judgment that *it is permissible to share food with one's kin*, or that *it is obligatory to share food with one's kin*. Alternatively, one would expect that a moral sense which disposes a person to view food sharing among kin as obligatory would tend to give that person a social and biological advantage over someone who views food sharing as forbidden. Similarly, a group or culture in which food sharing is forbidden is probably at a disadvantage compared to a culture that views food sharing as obligatory. In short, the content of moral judgments is plausibly thought to be subject to forces like natural and cultural selection, at least indirectly if not directly. Thus, the proper function of the moral sense will likely involve the disposition to produce moral judgments with fairly specific content.

Furthermore, as Richard Joyce (no moral realist, he) contends, our moral sense has probably been 'designed' to pick up on certain descriptive features of the world, but not others. Such descriptive features include, "examples of purposeful injury, the maintenance of reciprocal relations (fairness, cheating, etc.), social status, and a cluster of themes pertaining to bodies and bodily functions."[14] A bit more specifically, then, the proper function of the moral sense probably includes producing *X is good / X is right / X ought to be done* sorts of judgments in the presence of descriptive features of the world such as motives, actions, and the consequences thereof that are conducive to human physiological, psychological, or social well-being. Also, it probably has the proper function of producing "is bad / is wrong / ought not to be done" judgments in the presence of descriptive features inimical to human physiological, psychological, or social well-being. In sum, the proper function of the moral sense is to cause us to value the sorts of things that tend to make life go well for us physiologically, psychologically, and socially, and to disvalue the sorts of things that tend to make life go poorly for us physiologically, psychologically, or socially.[15]

PROPER-FUNCTION MORAL REALISM

What I want to suggest is that facts about the proper function of the moral sense can provide the truth conditions for at least some moral judgments. Likewise, facts about the proper function of the moral sense can ground the rightness or wrongness of actions as well as what sorts of considerations count as moral reasons for action. Call the resulting view *proper-function moral realism.* It is a form of realism because the proper function of the moral sense, and hence the truth conditions for at least some moral judgments, does not depend for its existence on what humans happen to have agreed to, or what humans would agree to if they were fully informed, or the like. Rather, as is the case with other biological entities that have a proper function, such as the human heart or lungs, facts about the proper function of the moral sense are what they are regardless of what anyone believes about them.

Proper-function moral realism offers the following account of the truth-conditions of moral judgments; what makes them true, which descriptive facts are part of the subvenient base upon which the rightness or wrongness of an action supervenes, and so forth: A token moral judgment M is correct

---

and Gibbard, who writes that the ''biological function [of normative judgements] is to govern our actions, beliefs, and emotions'' (1990), p. 110." quoted in Street (2006), p. 159. See also Joyce (2005), chapter four.

[14] Joyce (2005), p. 140.

[15] I cash out evolutionary benefit in terms of physiological, psychological, and social well being because I take it that having the ability to recognize when some course of action will help you and your conspecifics stay alive (and when an action will not), and possessing the motivation to avoid physiologically, psychologically, or sociologically detrimental dispositions or actions is a significant evolutionary advantage.

(true) in some particular situation or circumstance C when it is the proper function of the moral sense to produce M in C. To recycle an earlier illustration, suppose that someone is in a situation where food is relatively hard to come by. On any given day, one might fail to find anything to eat. However, on this particular day one has happened upon enough food to last more than a day, but one is faced with the decision of whether to share the food with someone who can be trusted. Since historically it has been advantageous for humans to share food with those who are trustworthy—perhaps especially so in times of scarcity—it is plausibly the case that the proper function of the moral sense in a circumstance like this would be to motivate foo d sharing via the judgment that it is right—perhaps even obligatory—to share one's food with those who are trustworthy. According to proper-function moral realism, this grounds (at least in part) the fact that it is right in that situation to share one's food with those who are trustworthy.

Typically, insofar as proper-function attributions entail the possibility of malfunction, the yield a sort of normativity, but only a very minimal sort. One might grant that the human moral sense has the rather complex, biologically proper function that I have described while denying that the proper function of the moral sense yields any sort of prescriptivity when it comes to human motives or actions. Rather, one might think that prescriptivity involves (if not requires) *objective reasons* for action. Proper-function moral realism can accommodate this line of thinking. Recall above the claim that the proper function of the moral sense includes tracking and using certain descriptive features of the world in the production of moral judgments; ostensibly, facts about the consequences of an action, whether one would want others to act in a similar manner, and so forth. On this view, then, that a descriptive fact of type D obtains in a particular situation counts as a moral reason for person P to do action X in circumstance C just in case it is the proper function of the moral sense to use D as one of its inputs in creating an "X is right" judgment in C. If, as I have argued thus far, the proper function of the moral sense is to cause us to value the sorts of things that tend to make life go well for us physiologically, psychologically, and socially, and to disvalue the sorts of things that tend to make life go poorly for us physiologically, psychologically, or socially, then descriptive facts about what sorts of motives or courses of action would affect physiological, psychological, or social well-being would also have moral weight to them.

Facts about the proper function of the moral sense, combined with the above account of what sorts of descriptive features count as moral reasons for action, suggest a particular account of moral supervenience. Where ϕ refers to some specific action, suppose that X ought morally to ϕ in C. What are the descriptive base entities upon which this moral fact supervenes? Proper-function moral realism holds that moral facts supervene on facts about the proper function of the moral sense and facts about the circumstances in which one finds oneself. Expressed schematically, proper-function moral realism holds that if X ought morally to ϕ in C, it is because (i) X is human, (ii) it is the proper function of the human moral sense to value those dispositions and actions that tend to promote human physiological, psychological, and social well-being, and because (iii) ϕ-ing in circumstance C tends to promote human physiological, psychological, and social well-being.

PROPER-FUNCTION MORAL REALISM AND TWO RECENT CHALLENGES

With the main details of proper-function moral realism sketched out, it is time to consider how it handles some of the more prominent antirealist challenges; namely, Sharon Street's, "Darwinian Dilemma" for moral realists and Richard Joyce's, "evolutionary debunking of morality," based on a phenomenon that he calls *practical clout*. As Street poses it, the dilemma for realists is that there is no plausible way to explain the relationship between the forces of natural selection and the content of moral truth. On the one hand, the realist might deny that there is any relationship; i.e., the content of moral truth is independent of evolutionary pressures on human moral evaluative tendencies. If she goes this route, then Street objects that, given the strong influence of natural selection on our evaluative

attitudes, the realist ought to conclude that these evolutionary forces can only exert a distorting influence on our evaluative judgments, and therefore most of our evaluative judgments are probably off track.[16] On the other hand, if the realist claims that there *is* some relation between the workings of natural selection and the content of moral truths, Street contends that whatever explanation the realist offers will likely be explanatorily inferior to an antirealist account that Street calls the *adaptive link account*. The realist account that Street considers is what she calls the *tracking account*. Briefly, the tracking account claims that making certain evaluative judgments contributed to reproductive success because those evaluative judgments are true, and because it proved advantageous to grasp evaluative truths. By contrast, the adaptive-link account claims that we tend to make certain judgments (e.g., that *the fact that someone has harmed one is a reason to help them in return*) because those judgments contribute to reproductive success, and this because making these sorts of judgments tended to get our ancestors to respond to their circumstances with behavior that itself promoted reproductive success. Street contends that there are at least three respects in which the adaptive link account is superior to the tracking account. First, because the tracking account posits something that the adaptive-link account does not (namely, mind-independent moral facts), the tracking account is less parsimonious than the adaptive-link account. Second, the adaptive-link account is clearer than the tracking account in that the tracking account does not explain the connection between evaluative fitness and the recognition of certain evaluative truths. That is, the realist does not explain why it should be evolutionarily advantageous to recognize moral truths in the first place. Third, the tracking account has nothing informative to say about all the normative judgments that human beings could make but don't; e.g., that *the fact that someone has harmed one is a reason to shun that person or retaliate*.

Let us now consider how proper-function moral realism addresses Street's challenges. According to proper-function moral realism, the relationship between the forces of natural selection and the content of moral truth is as follows: the content of moral truth is determined by the proper function of the moral sense. In turn, the proper function of the moral sense is determined by facts about what sorts of actions (and thereby judgments) tend to promote the long-term physiological, psychological, and social well-being of humans.[17] The proper-function realist account thereby combines elements of the tracking account and adaptive-link accounts. Like the tracking account, the proper-function account claims that it is evolutionarily advantageous to grasp certain facts; in this case, facts about what sorts of actions and moral judgments tend to promote the long-term physiological, psychological, and social well-being of humans. Unlike the tracking account, however, the proper-function account is not committed to the view that the moral sense conveys reproductive fitness by tracking moral facts *per se*. Rather, the moral sense conveys reproductive fitness by tracking what, via its possessing a proper function, are moral reasons for action. Like the adaptive-link account, the proper-function account holds that we tend to make certain judgments and not others because certain judgments and the actions they motivate tend to convey a reproductive advantage on the people who make them. Finally, it is not clear that the adaptive-link account is an account of the relationship

[16] Street (2006), pp. 121-122. The realist might object that our ability to reflect on our evaluative attitudes and judgments can counteract these *ex hypothesi* distorting evolutionary forces. Street claims that the problem with this objection is that, on the one hand, when we engage in rational reflection, we must always have some starting point; namely, some starting "fund of evaluative judgments," that we use to evaluate other claims. On the other hand, the fund of evaluative judgments that we humans start with has already been contaminated by illegitimate (i.e., evolutionary) influences. Thus, the tools of rational reflection are contaminated form the start, and are therefore not reliable tools to use to get at moral truth. (Street [2006], p. 124.)

[17] Philosophers disagree about whether the proper function of a mechanism or trait depends on what it has *historically* done to benefit the organisms that possess it, or whether the proper function of a trait or mechanism depends on what it does *now* to help ensure the survival and reproduction of the organisms that have it. For a survey of some of the major contours of this debate, see Buller (1999). While in this essay I proceed as though some form of etiological view is probably correct, nothing about proper-function moral realism *per se* hinges on this debate.

between evolutionary pressures and the content of moral truth at all. By itself, the adaptive-link account says nothing at all about moral truth. Rather, it only seeks to explain why we make certain moral judgments and not others; it does not specify what would make some of those judgments *objectively true*. By contrast, the proper-function account *does* indicate a link between evolutionary pressure and the content of moral truth, thus giving it an advantage over the adaptive-link account.

Regarding Street's three criteria of parsimony, clarity, and explaining why some logically possible moral judgments are rarely if ever made, the proper-function view seems in a good position to go toe-to-toe with the adaptive-link account. Consider parsimony. On the proper-function view, moral facts are grounded in otherwise ordinary, natural facts. These facts were in some ways shaped by selection pressures, but also by brute facts about human physiology. So, while the proper-function account necessarily includes entities that the adaptive-link account lacks; namely, moral facts, this does not appear to be any more problematic than proper functions elsewhere in biology. Regarding clarity, the proper-function account gives a clear explanation for why it should be evolutionarily advantageous to have true moral beliefs: Having true moral beliefs entails having true beliefs about what sorts of motives and actions tend to make life go well physiologically, psychologically, and socially for the people that have them. Furthermore, insofar as moral beliefs tend to motivate action, having true moral beliefs will tend to motivate physiologically, psychologically, or socially beneficial actions more often than having substantially false moral beliefs. Finally, the proper-function account is just as good as the adaptive link account when it comes to explaining why some moral judgments are almost never made. For one thing, the proper-function account can agree with the adaptive-link account in holding that we tend not to make certain value judgments because, historically, people who valued motives and actions that are generally harmful have died off and those who valued beneficial motives and actions survived and left more offspring. Moreover, some moral judgments are almost never made because in most people, the moral sense tends to function properly to some degree, and because the proper functioning of the moral sense involves tending not to form judgments that would be radically harmful to the organisms that would form them. Further, if we assume that the principle of bivalence holds for adequately specified moral judgments, then the proper-function account can explain why certain judgments are false. Specifically, some moral judgments are false in a specific set of circumstances because their content does not correspond to the content of the judgments that it is the proper function of the moral sense to produce in those circumstances. For that matter, it cannot have been the proper function of the moral sense to generate judgments that, on balance, tend to be highly detrimental to the species. In sum, Street's "Darwinian dilemma" is no dilemma for the proper-function moral realist.

Whereas Street challenged the moral realist with a dilemma, Joyce's challenge to moral realism is more straightforward. At bottom, his claim is that the evolutionary history of our ability to make moral judgments undermines the rationality of our moral beliefs. How so? On the one hand, it seems to us that when we make moral judgments, we are responding to or tracking certain attributes of the world; namely, authoritative moral facts or principles. That is, the what-it-is-like aspect of moral judgments typically involves a sense that some actions morally *must* be done, and other actions morally *must not* be done. Similarly, some moral rules, such as *do not torture infants for fun* seem to be morally inescapable; that is, they apply to us independently of our desires or interests, and independently of whether we recognize them at all. Further, moral rules seem to have a genuine claim to our allegiance in the sense that, were we to reason correctly, we would want to abide by them.[18] Unfortunately, while our moral faculties make it seem to us as though we are responding to some inescapable, authoritative, in-the-world rules, it is unlikely that any natural facts could match the above description. Rather, much like his predecessor J.L. Mackie, who claimed that there are no objectively prescriptive facts, Joyce contends that no natural fact or set of natural facts could possess the

[18] Joyce (2005) pp. 195-196.

inescapability and authority that at least some moral principles or rules would have to possess if moral realism were true.[19] In short, evolution has left us prone to regularly viewing the world as possessing features it does not actually have; all our moral beliefs are the result of unreliable, moral-belief-producing mechanisms.

The proper-function moral realist can address Joyce's concerns in the following way: Recall that according to proper-function moral realism, the facts about what one morally ought or ought not to do are grounded in the proper function of the moral sense, in conjunction with facts about what courses of action tend to make life go well for humans physiologically, psychologically, or socially. Thus, moral rules—at least some of them, anyway—are inescapable insofar as we cannot change the proper function of the moral sense any more than we can change the proper function of our heart, lungs, or brain. By and large, any human with the capacity to make moral judgments is thereby under the purview of morality. With regard to moral authority, recall that Joyce's claim that a rule is authoritative if, were one to reason correctly, one would want to abide by it. The proper-function moral realism lends itself to the following account of the correctness conditions for moral reasoning: the actions that tend to be fitness enhancing for the species are precisely those that are such that we would want to do them if we were to reason correctly. In other words, were one to engage in correct moral reasoning, one would value the sorts of actions and motives that tend to bring physiological, psychological, or social benefit, and one would not value the sorts of actions and motives that tend to bring about physiological, psychological, or social harm. On this approach, reasoning correctly would involve using whatever tools are most reliable for discovering which sorts of actions and motives tend to benefit (or harm) humans. Likewise, it would involve cultivating whatever feelings tend to motivate beneficial actions and discourage harmful ones. *Pace* Joyce, scientific discovery could, in theory, settle moral disputes.[20]

CONCLUSION

Proper-function moral realism has at least four things going for it. First, is broadly in line with a reasonable explanation of human origins, including the origin of our ability to moralize. Secondly, it offers an explanation for why certain descriptive facts count as moral and others do not. Third, since it relies only on putatively natural concepts and entities like proper functions, reproductive fitness, physiological well-being, and so on, it requires no metaphysically "queer" entities. Fourth, it can straightforwardly answer some of the more powerful objections to moral realism that are currently on offer. For at least these reasons, it merits consideration as a serious metaethical view.

BIBLIOGRAPHY

Blackburn, Simon (1993). *Essays in Quasi-Realism* (Oxford University Press)
Buller, David J. (1999). *Function, Selection, and Design* (SUNY Press)
Casebeer, William and Patricia Churchland (2003). "The Neural Mechanisms of Moral Cognition," *Biology and Philosophy* 18: pp. 169-194.
Gibbard, Allan (1990). *Wise Choices, Apt Feelings* (Harvard University Press)
Griffiths, Paul (1993). "Functional Analysis and Proper Functions," *British Journal for the Philosophy of Science* (44): pp. 409-422
Joyce, Richard (2005). *The Evolution of Morality* (MIT Press)
Mackie, John (1977). *Ethics: Inventing Right and Wrong* (Oxford University Press)
Millikan, Ruth (1989). "In Defense of Proper Functions," *Philosophy of Science* 56 (June) pp.288-302.
Neander, Karen (1991). "Functions as Selected Effects: the Conceptual Analyst's

[19] Cf. Mackie (1977), chapter one.
[20] Compare Joyce (2005), chapter five.

Defense," *Philosophy of Science* 58 (2):168-184.
Street, Sharon (2006). "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies* 127 (1):109-166.
Woodfield, Andrew (1976). *Teleology* (Cambridge University Press)

Stephen Davey

University of Texas at Austin

Stephen Davey is a graduate student at the University of Texas at Austin, where he works mainly in the areas of metaethics and the theory of action. He is especially interested in marrying normative realism with an Anscombean conception of intentional action. From time to time, he also finds himself thinking about the relations between reasons for action and other types of reasons, and about the nature of explanation.

"The Problem With (Quasi-Realist) Expressivism"

Abstract: *This is an essay about how expressivism about the normative understands the notion of a reason for acting. I claim that no adequate understanding of this notion is available to expressivism and that this is because of the basic strategy of the view: that it attempts to understand the meaning of normative utterances in terms of the mental states expressed by those utterances. Or, as I prefer to say, it attempts to understand what it is to identify a normative feature in terms of what one does when one treats some object as having a normative feature. I present a counterexample designed to show that some normative judgements put identification and treatment at odds and therefore require a distinct understanding of each.*

Introduction

This is an essay about how expressivism about the normative understands the notion of a reason for acting. The thesis is that no adequate understanding of this notion is available to expressivism and that this is because of the basic strategy of the view: that it attempts to understand the meaning of normative utterances in terms of the mental states expressed by those utterances. Or, as I prefer to say, it attempts to understand what it is to *identify* a normative feature in terms of what one does when one *treats* some object as having a normative feature. I present a counterexample designed to show that some normative judgements put identification and treatment at odds and therefore require a distinct understanding of each. For ease of presentation, I focus on Allan Gibbard's version of expressivism, which describes the mental states expressed by normative utterances as planning states, but the criticism is perfectly general. The difficulty that my counterexample reveals is due to the core expressivist strategy of understanding identification in terms of treatment, and not from any of the details of the planning account in particular.

In the next two sections I construct the dialectic. I describe the basic features of the expressivist strategy that I go on to argue are the root of the problem, and I lay out the elements of the quasi-realist move that I go on to argue are more of a misdirection from, and less of a solution to, this problem. This is all pretty cursory. I do not take myself to be engaged in anything like a thorough discussion of the merits and varieties of expressivism; I am only interested in bringing to light those features of the target view with which my objection makes contact and the location in the broader discussion where that contact takes place. In the final section, which is the meat of the essay, I give the objection and consider and respond to replies on behalf of the expressivist.

The Expressivist Strategy

Moral expressivism is supposedly a family of positions in metaethics, though one might be tempted to view it more as a revision of, or reaction to, metaethics.[21] Metaethical questions, or more broadly, meta*normative* questions, concern the meaning of normative language and the basis for the

[21] This may sound tendentious, but I do not mean to endorse this view here – I am merely setting the stage.

truth (if there is any to be found) of normative judgements and inquiries. Among other things, metaethicists are interested in defining normative terms in order to understand the meaning of sentences that employ them. Rather than giving 'straight' definitions, however, expressivism shifts the focus of the inquiry and seeks to understand normative utterances in terms of the mental states that they express.[22] One motivation for this shift in approach (the one that Alan Gibbard emphasizes[23]) is a suspicion that any plausible straight definition will employ normative language in the definiens that will itself require explication. Were this the case, a straight definition strategy might proceed by positing irreducibly normative, non-natural properties to supply the content of normative utterances. But one might not be satisfied with this type of explanation if one found these non-natural properties just as mysterious as the normative language they were invoked to explain. Skeptical of the possibility of giving a straight definition of the normative in non-normative terms, the non-naturalist might doubt that the best theory really *could* tell us anything more. But the expressivist will insist that her approach *does* tell us something illuminating about our normative experience, and that we ought not to think less of it simply because what it tells us is in some way less direct than we might have anticipated (especially considering the lack of explanatory promise in the alternative). Gibbard makes just such a claim in explaining his understanding of the concept of a normative reason.[24]

Now, there are some things to be said for resisting this expressivist strategy and sticking with the ordinary 'descriptivist' understanding of the meaning of normative judgements and utterances that the non-naturalist favours.[25] First, one might think that the expressivist does not provide a better description than the descriptivist since, in changing the nature of the inquiry, the expressivist searches for a different, non-competing explanation from the descriptivist. Second, in making normative claims, we really do take ourselves to be saying things that are truth apt, and furthermore true, and further still, true independent of our endorsement of them, but none of this is the case according to expressivism. Expressivism does not seem to respect the phenomenology of our normative experience.[26]

In response to these concerns, some expressivists, Gibbard included, have built their theories around an emphasis on the compatibility of expressivism with these aspects of the phenomenology of our normative (and metanormative) experience. The goal is to understand normative judgements and language naturalistically, in terms of the mental states they express, while still holding on to all the trappings of descriptivism.[27] If this so-called 'quasi-realist' expressivism can make sense of everything the descriptivist insists on, and can do so in a way that respects these aspects of our normative experience, then it preempts the descriptivist's objections and gains a dialectical advantage.

---

[22] Allan Gibbard, *Thinking How to Live*, (Cambridge: Harvard University Press, 2003), p. 6. Cf. Allan Gibbard, *Wise Choices, Apt Feeling*, (Cambridge: Harvard University Press, 1990) p. 8: "The analysis is not directly of what it is for something to *be* rational, but of what it is for someone to *judge* that something is rational" (original italics).

[23] Gibbard, *Thinking*, p. 6.

[24] Ibid, pp. 185, 190-91.

[25] I will be referring to the expressivist's main target alternately as 'descriptivism' and 'non-naturalism' depending on what seems most fitting at the moment. Sometimes it will be the non-naturalist's descriptivism in particular to which he is reacting, and other times it will be her non-naturalism.

[26] And of course there is the Frege-Geach problem. Though this is, of course, an important part of the motivation for going quasi-realist, it will not feature in my discussion.

[27] I'm borrowing this snappy phrase from Richard Joyce, "Moral Anti-Realism", *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2009/entries/moral-anti-realism/>.

The Quasi-Realist Move

The meaning of normative language (and the content of normative judgements) can be understood, for Gibbard, on a model of possible worlds.[28] The states of mind that I express with my normative utterances are planning states, and planning can be understood as allowing or ruling out certain activities for certain contingencies. In uttering "It was wrong for Edward to disrespect Linda that way," I express my ruling out the option of disrespecting Linda for the contingency of being in Edward's shoes. In uttering "I ought to go down to the soup kitchen tonight," I decide on going down to the soup kitchen for the contingency of being me later tonight (which is to say, I rule out not doing so for that contingency). Planning consists in making commitments that restrict the set of possible combinations of circumstances and actions, and the content of an utterance that expresses a plan is fixed by the possible combinations that it allows and rules out.

Giving the content of normative utterances in this way makes available a logical framework that parallels truth-functionality for descriptive language.[29] In doing so, Gibbard gets part of the way to a quasi-realism that appears to enjoy the benefit sketched in the last section.[30] Two things are missing. First, the thought that metanormative judgements – even the theory-laden judgements of the non-naturalist that might *appear* to speak directly against expressivism – can also be given an expressivist gloss. For example, claims about attitude-independence might be understood as plans for which the relevant sets of possible worlds are not constrained by any particular set of attitudes.[31] If we take this suggestion along with what has been said so far about ordinary normative judgements, it would seem that the expressivist has something to say about the meaning of any utterance that employs the normative concepts that have otherwise seemed mysterious. Now, if we favour a minimalist notion of truth such that we understand the meaning of "p is true" as nothing over and above the meaning of "p", and if the expressivist is perfectly prepared to say something about the meaning of any normative (or metanormative) utterance (or judgement), then there seems to be nothing more to say about *the truth of* any normative (or metanormative) utterance (or judgement) than what the expressivist is prepared to speak to. If all of this is right, then there is a sense in which the QREist can grant the non-naturalist anything she likes; the non-naturalist has no claim against the QREist that she is leaving something unaddressed.

I began the last section by suggesting that expressivism might be best thought of as a response to metaethics rather than a position in metaethics. It is worth pointing out in concluding this section that the quasi-realist component makes QRE seem more like metaethics again since it accepts a characterization of the project as an attempt to understand the meaning of normative language. It just understands "the meaning of normative language" in a non-standard way. So QRE is set up to be more of a contender with non-naturalism than one might initially have thought. If Gibbard's view can be made to work, he can say, in a sense, everything that a non-naturalist can say, and can do so without appeal to non-natural facts or properties. Furthermore, by understanding normative facts in terms of psychological facts, he can provide more in the way of explanation than the non-naturalist who can do little more than rely on brute normative facts.

In the next section, however, I will argue that the quasi-realist move does not succeed; it merely conceals the problem. Because the general expressivist strategy involves understanding the identification of some feature as a reason to act in terms of what one does when one treats that feature

[28] Gibbard, *Thinking*, Ch. 3, esp. pp. 46-7, 58-9.
[29] Ibid, p. 58.
[30] As well as other benefits not mentioned here, such as an explanation of the apparently tight relation between ethical judgements and motivations. Recall the disclaimer of the introduction.
[31] Ibid, p. 186.

as a reason to act, it leaves some types of normative judgements and utterances unexplained, namely those an understanding of which requires a notion of identification as distinct from treatment.

Identification and Treatment

There is a difference between (A) planning not to grant any weight to certain considerations in one's planning, and (B) planning not to grant certain of one's reasons any weight in one's planning. The former, and not the latter, can be done precisely because the relevant considerations provide no reason at all (though perhaps one might be tempted to treat them as though they do, and so one might plan in this way as a sort of self-imposed discipline). One might undertake the latter, but not the former, in setting out to do wrong, where one knows that one will not "go through with it" if one considers the moral landscape fully and allows one's conscience to take over. My objection is that this difference cannot be captured in expressivist terms.

Consider two cases. First, a member of a hiring committee is considering various candidates for the same job. He knows that considerations of race, gender, sexual orientation, political affiliation, etc. do not provide him with reasons one way or the other concerning whom to hire, but he also knows that the sort of biases that might influence him often operate subconsciously and in people who do not appear, at least, to be prejudiced. Having no delusions about being a special case, immune to the findings of social psychologists, he takes extra measures to ensure that these irrelevant considerations have a minimal impact on his decisions.

Second, a president has had to make a difficult decision. Despite what she takes to be very strong moral reasons to the contrary, she plans to order the covert assassination of another world leader. She judges that ordering the assassination is what she must do, but nonetheless she continues to feel the pull of the opposing considerations – they keep her up at night, she second-guesses herself, she worries about the guilt she might feel afterward. Furthermore, she knows that when the time comes she will have a very short window of opportunity to order the assassination, and that she is very likely to balk if she continues to give her anti-assassination reasons due consideration. She cannot allow that to happen, so she plans to ignore them. She plans not to attend to them, and so not to grant them any weight in her deliberations when the time comes to act.

The difference between these two cases is that the president, but not the committee member, identifies the considerations that she plans to ignore *as reasons*, that is, as counting in favour of the course of action she plans to avoid. But this characterization of the difference between the two cases is not available to the expressivist – it is just the sort of "straight" treatment of a normative concept that expressivism is in the business of avoiding. So how would an expressivist understand this added feature of the president's case so as to distinguish what is going on there from what is going on with the committee member? The expressivist understands the notion of a reason in terms of the mental state that is expressed by calling something a reason. In Gibbard's terms, we are planning to weigh it in favour of acting in a certain way.[32] But that does not look like an accurate characterization of the president's plans. On the contrary, her plan is precisely *not* to grant weight to the considerations she takes to be reasons not to order to assassination. If expressivism is to accommodate the difference between the president's and the committee member's cases, then, the feature of the analysis that accounts for that difference will have to be something other than the president's plan to weigh those considerations in her deliberations when the time comes.

Before going on to consider the other options for expressivism, I think it is worth taking a moment to say something more about what the president's case is not. First of all, the president is not judging irrationally: she is not judging both that certain considerations do and do not speak against

[32] *Thinking*, Ch. 9, especially pp.189-91. NB: if the contingency for which we treat R as a reason to X is our current situation, then we might just say that treating R as a reason to X amounts to weighing R in favour of Xing. This 'weighing' is to be understood as a purely psychological activity that could be mimicked by a robot.

ordering the assassination, nor is she planning both to weigh and not to weigh the same considerations against ordering the assassination.[33] Call T^1 the time at which she judges that she has good reason to ignore her anti-assassination reasons, and T^2 the time in the near future for which she plans to ignore them – the time at which she will have a brief opportunity to give the order, or perhaps just before then. At T^1 she recognizes the force of the anti-assassination considerations, but she plans to be in a state, once T^2 rolls around, such that she does not recognize their force (indeed, such that she does not attend to them at all). So, at T^2, if things go as she plans, she will not recognize the force of the anti-assassination considerations, nor will she be engaged in any planning with respect to them.[34] There is nothing in the case that suggests, either at T^1 or at T^2, that she is being, or planning to be, irrational.

Secondly, hers is not just a straightforward case of one set of reasons outweighing another set. Expressivism has something to say about that type of case: it is a matter of planning to grant some weight to the second set, but more weight to the first set, such that the first set wins out (or something along these lines). The president's case is more complicated for expressivism because she does not merely plan to grant some, but ultimately not enough, weight to her anti-assassination reasons; she plans to ignore them.

What are the options, then, for understanding the president's judgement that while certain considerations do count in favour of aborting the assassination, other considerations count decisively *against* weighing them in favour of acting in that way?

> *Plan*: Prez says "I have a reason to ignore certain of my reasons" and in so doing expresses a planning state, namely

(1) Prez's plan to weigh some consideration toward ignoring other considerations in Prez's deliberations in T^2

(2) Prez's plan to weigh some consideration toward ignoring other considerations *to which Prez would otherwise be inclined to grant some weight* in Prez's deliberations in T^2

I mention (1) as a contrast for the remaining options, just as the committee member's case provided a contrast for Prez's. (1) fails to identify the considerations which Prez plans to ignore as reasons, and so resembles (A) rather than (B). The italicized addition in (2) is an attempt to specify that they are reasons, and it is supposed to be in line with an expressivist understanding of reasons. Were *Plan* not operating, Prez's treating those considerations as reasons would amount to her planning to grant them some weight in her planning, so one might expect the added feature in Prez's case to be cashed out in those terms. But this is not in line with an expressivist understanding of reasons because it characterizes reasons descriptively.

The addition in (2) characterizes a reason as something the agent would be inclined to weigh toward an action. This is not what expressivism tells us about reasons. Expressivism seeks to understand the notion of a reason in terms of the planning states that are expressed by statements about reasons.[35] But, first, identifying some consideration as one I would be inclined to weigh in favour of some action is not planning. It is not a matter of settling on giving it weight now, nor is it a matter of committing (or merely allowing) myself to give it some weight for some hypothetical situation.

---

[33] The reader might still want to characterize the president's plans as irrational in some other sense. I don't think they are, but we needn't worry about that at the moment, so long as it is agreed that they are not irrational in this straightforward sense.

[34] Really, it doesn't matter whether things go as planned once T^2 rolls around. The judgement that is up for expressivist translation takes place at T^1, so that is the time at which it is important that we view the president as rational.

[35] Furthermore, it is supposed to do this in a "way we can describe without helping ourselves to the notion of a reason" *Thinking*, p. 190.

Rather, it is purely a description of that consideration as it relates to certain of my dispositions. Second, if the addition in (2) is to be understood as characterizing "the state of mind, in effect, of *believing* it to be a reason"[36] then it is just wrong. When we believe something is a reason, our belief is not about our dispositions, even if we are disposed to view as reasons the same things we are disposed to treat as reasons. The only other option I can think of is this:[37]

(3) Prez's plans, (i) to weigh some consideration toward resisting granting weight to other considerations of type *R* in Prez's deliberations in T^2, and (ii) to grant weight to considerations of type *R* in Prez's deliberations in circumstances other than T^2.

My first reply to this option is to ask which other circumstances feature in the second plan. It cannot be *all* other circumstances. Situations adequately similar to Prez's will also call for *R*-type reasons to be ignored (at least by her lights), and there will be some cases where *R*-type considerations do not provide reasons in the first place.[38] This strategy would require some mechanism for distinguishing the present case from others with respect to just those features that incline Prez to ignore her reasons while, at the same time, maintaining just those similarities between cases that give the reasons their status in T^2. I cannot see how this can be done expressivistically.[39]

Perhaps more to the point, however, is that even if this option does distinguish between how the agent views the relevant considerations in (A) from how the agent views the relevant considerations in (B), it does so in the wrong place. The addition in (3) is made not as a part of Prez's current plan, but as a separate plan for a separate contingency, and so excludes it from Prez's deliberations as she is confronted by the practical problem of T^2. But it is precisely the demands of the current practical problem that call for the distinction to be drawn – it is precisely because race does not count *in the committee member's situation* that he ignores it, and it is precisely because certain ethical considerations do speak against *the assassination Prez is considering* that she ignores them. Whether similar considerations would count (or whether either agent would be inclined to weigh them as such) in other circumstances is irrelevant to the reasons each has for ignoring the considerations in question in her respective actual circumstances. Whatever reasons there are for Prez not to order the assassination: whether these types of considerations would almost always make other assassination attempts wrong, or whether hers is one of the very few imaginable scenarios in which they would matter at all, it would remain the case that their counting *in her circumstances* contributes to the reasons *in her circumstances* for ignoring them (or so, at least, she judges). But this cannot be captured on (3). The normative status of *R*-type considerations – for Prez's circumstances or for any other – is left out of her plan for T^2. The only mention of their status as reasons is, first, a matter of their status in other contingencies and, second, isolated to her planning for those contingencies. If there is another candidate for characterizing the difference between (A) and (B) in planning terms, I cannot see it.[40]

Here, I think, is the general problem that the president's case presents for expressivism. The expressivist strategy has us understand the normative notion of a reason in terms of the mental state that is expressed by calling something a reason. When we are in this mental state, let's say we are

---

[36] Gibbard, *Thinking*, pp. 190-1, italics added.
[37] Thanks to Brian Cutter for suggesting this possibility, and for helpful discussion.
[38] For example, suppose R includes worries about slippery slopes in foreign policy. We can imagine a case where all significant players on the world stage are in agreement that the circumstances are unique and no dangerous precedent is at issue.
[39] In fact, I am not sure that it can be done in any other way either. For a compelling case, see Jonathan Dancy's discussion of reasons holism in his *Ethics Without Principles*, (Oxford: Clarendon Press, 2004).
[40] I have left out what I take to be too implausible to count as an option: that *Plan* expresses a plan to both weigh and not weigh those considerations in deliberating in T^2.

treating the relevant consideration as a reason, rather than *identifying* it as one in the sense that would appeal to a descriptivist, which might amount to judging or saying of it that it counts, that it bears the favouring relation to some type of action, or something along these lines. I take it that we all have a decent pre-philosophical grasp on the notions of identification and treatment and the differences between them, but if the reader wants to hear a little more before going on with the argument, I can offer at least this much: to identify some consideration as a reason is to say something of it (that it has some property or stands in some relation), whereas to treat some consideration as a reason is to take a stance toward it. Taking a stance toward something need not involve any sort of attribution, and in the context of expressivism it necessarily won't. It might instead be a matter of making use of, being moved by, making commitments with respect to, etc. Expressivism, then, has us shift focus away from *identifying* considerations as reasons (and describing them as such) and toward *treating* considerations as reasons (in planning to weigh them in our deliberations[41]). But examples can be designed to force the spotlight back onto identification as distinct from treatment. Prez cannot be construed as treating the relevant considerations as reasons because (and this is the key feature of her case) her plan is precisely to treat them as though they are *not* reasons. But this does nothing to change the fact that she identifies them as reasons; it couldn't, because their counting against the assassination is just the feature by which she picks them out as needing to be ignored in the first place. If she no longer saw them as counting one way or the other, she would not be concerned that they might engage her conscience and stall the operation. If understanding language that (prior to being convinced of expressivism) we would ordinarily think of as doing the work of identification requires that instead we think of the utterance as indicative of treatment, how can we make sense of cases that require a distinct understanding of both?

If one took the achievements of the quasi-realist move – that it allows expressivism to (i) account for even those metanormative judgements that appear, at first, to be out of reach, and (ii) go on making (minimalist) attributions of truth – to put QRE on level ground with descriptivism, one might find the purported explanatory advantage of QRE to be a compelling tie-breaker. But as soon as we realize that identification and treatment are genuinely distinct and that expressivism is ill equipped to make the distinction, that appearance dissolves. While expressivism may have something to say after all about the metanormative judgements that appeared, at first, to be out of reach, this does not address the question whether what it does say (in these and all other cases) is adequate. The theory may indeed be versatile enough to say *something* about what's going on with someone who says, for example, "that would be a reason for me to help him no matter what my beliefs and desires," but the one thing it cannot say is that she identifies that feature as such a reason. And this is something it must not leave out if it is to adequately capture that notion, or so my objection is meant to show. The persisting problem is that recognizing of something that it matters is not the same thing as treating it as though it matters. No mechanism designed to give normative claims the mere appearance of ordinary descriptive ones addresses this problem so long as that appearance ultimately bottoms out expressivistically, in terms of treatment.[42]

[41] Whatever alternative versions have in the place of Gibbard's planning states can fill in these parentheses just as well.

[42] I owe thanks to Jonathan Dancy, Alan Gibbard, Brian Cutter, and audiences and commentators at Gonzaga University, Brown University, SUNY Albany, and the 2012 APA Central Division meeting for their feedback on earlier drafts of this paper.

Arudra Burra

UCLA

Arudra Burra is a Postdoctoral Scholar in the Law and Philosophy Program at UCLA. He studied philosophy, mathematics, and computer science at Brandeis University (2000), earned a JD from the Yale Law School (2007), and a PhD in Philosophy from Princeton University (2011), for a dissertation entitled "Coercion, Deception, Consent: Essays in Moral Explanation." In addition to his academic work, Arudra was involved for many years with the "Right to Food" Campaign, an informal network of organizations and individuals committed to the goal of ensuring food security in India. He also been an active member of the Law and Social Science Research Network (LASSNet), a network of activists and academics with shared or overlapping interests relating to the law in South Asia.

"Coercion and Moral Explanation"

Abstract: *A dominant view of coercion takes it to essentially involve threatening another with an unwelcome consequence if they refuse to comply with what one demands. I argue, on a variety of grounds, that no such "threat-based" account of coercion can explain why coercive acts are wrongful. In fact, I claim that no general theory of coercion can bear this explanatory burden. While coercive acts share a certain structure, at a deeper level the principles governing which coercive acts are wrongful are quite diverse. To discover what they are requires substantive moral reasoning about acceptable ways of getting people to do things in the domain in question (e.g. the domain of property transactions or the domain of sexual relations). So it is a mistake to think that coercive acts are wrongful because coercive; rather the reverse. A similar strategy is used to side-step, or rather embrace, the long-standing "paradox of blackmail."*

I

In a well-known paper on the right to privacy, Judith Jarvis Thomson (1975) argues that every right in the "right to privacy" cluster is also in some other cluster of rights, e.g. the right to property, or the right to be free of annoyance; there are no rights to privacy "over and above" these more particular rights.[43] There is then, she claims,

> no need to find the that-which-is-in-common to all rights in the right to privacy cluster and no need to settle disputes about its boundaries. For if I am right, the right to privacy is 'derivative' in this sense: it is possible to explain in the case of each right in the cluster how come we have it without ever once mentioning the right to privacy. Indeed the wrongness of every violation of the right to privacy can be explained without ever mentioning it. (Thomson 1975, 313).

I wish to propose a similar, "reductionist" view of coercion, on which the wrongfulness of every act of coercion is to be explained by reference to a more particular moral principle governing how to get people to do things in the domain or context in question. On this view of coercion, wrongfully coercive acts share a common structure – they involve getting someone to do something in a way that is (other things equal) wrongful or improper.[44] But the fact that an act is coercive plays no independent role in the explanation of why it is wrong; we must explain why these acts are wrong by appeal to moral principles which do not themselves invoke the notion of coercion.

---

[43] Thanks to Be Birchall for pointing me towards this article.

[44] Here I follow hints in Frankfurt (1988) and Scanlon (2008).

The argument is somewhat indirect. I start by discussing the idea that coercion is constituted by threats to inflict penalties on another in if they do not comply with what one demands. I argue at some length (in §§II-IV) that this view of what coercion is cannot adequately account for the idea that coercion is *pro tanto* wrong. I claim that the connection between threats and coercion is contingent and derivative; threats are only a *means* of coercion, but they do not always function as means. In §V I suggest that the reductionist thesis offers a way to understand the so-called "paradox of blackmail," and in VI I extend the point to coercion more generally. Here I also suggest that we liberalize our notion of how one can go about coercing another: I suggest that offers, lies, forms of psychological manipulation, and physical forcings all have an equal claim to the title 'coercive,' at least in principle. In §VII I embrace a more eliminitavist conclusion regarding the role played by the fact that an act is coercive in an explanation of why it is *pro tanto* wrongful.

These claims sound more revisionary than they are. I think we are, in practice if not in theory, already reductionists about coercion. What I have tried to do is simply provide an alternative characterization of phenomena that are already familiar, and widely discussed in the philosophical literature on coercion. What I hope is that this alternative way of looking at things unifies these phenomena in more natural and fruitful ways.

II

Many philosophers have thought that coercion essentially involves threatening another with unwelcome consequences if they do not do what one tells them to do. Here they follow, by and large, a tradition established by Robert Nozick in a seminal paper (Nozick 1969).[45] They have subscribed to something like the following principle:

> **C.** P coerces Q into Fing just in case P threatens Q with a penalty which renders not Fing, with the penalty, substantially less eligible than it was without the penalty; and Q Fs in order to avoid the threatened penalty.

C requires a comparison, from the point of view of Q, of two courses of action: not-Φing and incurring a penalty A, and Φing without incurring the penalty; the penalty being such as to make not-Φing "substantially less eligible" than Φing. Q's having to make this choice is a consequence, in the paradigm case, of a certain speech-act of P's. This is the utterance of a conditional proposal of the following form: If (and only if) you refuse to Φ, I will bring it about that A.[46]

Coercion involves limiting or infringing another's freedom, and freedom might involve (at least in part) having meaningful choices or unburdened options. Threats remove a choice or burden an option, such as the option to keep one's money rather than giving it to the highwayman; in this way they limit another's freedom. So part of **C**'s appeal lies in the straightforward connection it seems to establish – via the notion of a threat – between coercion and the reduction of another's freedom.

Now **C** is a theory of coercion, not a theory of *morally problematic* coercion. But 'coercion' is hardly an idle wheel in moral theorizing: in the absence of special justification, it is wrong to coerce another; and when one wrongfully coerces another one seems to wrong them in a distinctive way. A theory of coercion had better tell us how this could be so.

It might be wondered whether **C** is up to the job. For consider cases such as the following:

---

[45] The most systematic exposition within this tradition – to which I am considerably indebted – is by Alan Wertheimer (1987). Of course, the idea that coercion involves "orders backed by threats" is hardly a new one; it figures, for instance, in Austin's "command theory" of law. Nor is the Nozick-based tradition the only one: Philip Pettit has urged in many writings (e.g. Pettit 2007) a "republican" conception of coercion which takes as central the notion of domination. I believe that this republican tradition of thinking about coercion has room for the view articulated here, but I do not discuss the connection. Frankfurt's (1988) account of coercion is also quite different from Nozick's, though it too places a great deal of importance on the notion of a threat. I am considerably indebted to this paper, though the connections with the view developed here may not be evident on the surface.

[46] The "only if" is often implicit: when the gunman threatens to kill you unless you hand over your money, there is the unspoken implication that he won't kill you if you do.

(1) If you don't come in to work on time I'll fire you from your job.

(2) If you don't give me a raise I'll join your competitor.

(3) If you don't fulfill your campaign promises I won't vote for you again.

(4) If you don't reduce your costs I'll take my business elsewhere.

(5) If you don't stop beating your wife I'll tell the cops.

Let me stipulate that these proposals ("hard bargains") hold out the prospect of rendering a course of action substantially less eligible than it would have been absent the proposal. Let me also stipulate that in the circumstances the utterance of these proposals is morally unproblematic, where I mean by this more than mere moral permissibility. It is not wrong *at all* to make these utterances in these contexts.

If hard bargains are threats and threats are coercive, and coercion is (other things equal) wrongful, we seem to have a problem, since hard bargains are not (even other things being equal) wrongful. So we have a choice. Either hard bargains aren't threats (in the morally relevant sense); hard bargains are threats but aren't coercive threats; or hard bargains are coercive threats, but aren't wrongfully coercive threats.[47]

I don't think much hangs on which of these options we choose, as long as we locate the relevant distinction somewhere, keeping in mind that we want to strike a balance between a notion of 'threat' that accords with out ordinary linguistic intuitions, and one that plays a fruitful role in moral theorizing. I suggest we count hard bargains as threats, and continue to regard coercion as *pro tanto* wrongful. This requires making a distinction between threats *simpliciter* and coercive threats. One will then have to amend **C** to require the condition – whatever it is – which threats must meet in order to be coercive.[48] That is the task of the following section.

III

Let us introduce some terminology. Suppose P makes Q a conditional proposal of the following form: 'If (and only if) you refuse to Φ, I will bring it about that A'. Call P here the "intervener," Q the "recipient," and the bringing about of A the intervener's "declared unilateral plan."[49] This is the act that the intervener would perform if the recipient refuses to Φ. The intervener's declared unilateral plan provides the recipient with an incentive to comply with her request. When the declared unilateral plan is a penalty, the proposal is a threat; when it is the 'mere' withholding of a benefit, it is an offer.

I will not be concerned here with the distinction between threats and offers, though a large part of the post-Nozick literature has been addressed to this very question. This is partly a matter of the terminological convention I adopted above, on which hard bargains count as genuine threats. Our question is how to distinguish threats *simpliciter* from (wrongfully) coercive threats, so that hard bargains don't count as wrongfully coercive.[50] It is natural to locate this distinction in what the intervener threatens to do if the recipient refuses to comply – that is, in the intervener's declared unilateral plan.

The suggestion is that what makes a threat coercive (hence impermissible) is that it is impermissible to carry out the intervener's declared unilateral plan. The thought can be made concrete

[47] A fourth option – they are wrongfully coercive, but certain features of the situation make it the case that they are not wrongful "all things considered." I reject this possibility because it's unclear what such additional features are and how they contribute to the all-things-considered judgment. I discuss similar issues further in "Deception and the Structure of Moral Principles."

[48] Here I part ways with Wertheimer (1987), who takes the relevant distinction to be between threats and offers, rather than between coercive and non-coercive threats. But this is largely a matter of terminology.

[49] Some of this terminology is borrowed from Scanlon (2008). The phrase "declared unilateral plan" is due to Haksar (1976).

[50] The parenthetical "wrongfully" is inserted only as a reminder: it is strictly speaking, redundant. For recall I am taking coercive threats to be *pro tanto* wrongful. (So, while permissible coercion is a genuine possibility, the permissibility of making coercive threats will depend upon further circumstances – such as a circumstance in which making a coercive threat is the only way to avert a very great harm).

in more than one way, and in this section I want to survey some of these ways. I will suggest, eventually, that while these "Inheritance Principles" draw the right distinctions, it is obscure just how the wrongfulness of making a coercive threat – a particular speech-act – is to be explained by the wrongfulness of a hypothetical act which the intervener may never have intended to perform, and which, when the threat succeeds in coercing the recipient, will not in fact have been performed.

Let us begin with the simplest version of such a principle, which one might call a principle of *unrestricted* inheritance:

> **Unrestricted Inheritance**: If it is impermissible to threaten to A (if another does not Φ), then that is because it is impermissible to A (independently of whether or not the other Φs).[51]

The principle of unrestricted inheritance seems to sort the cases correctly: the highwayman's threat is wrong because it would be wrong to shoot his victim if he does not hand over the money. And clearly it also excludes hard bargains: what makes it permissible for me threaten to quit unless I get a raise, or to fire you unless you come to work on time, is that it would in fact be permissible for me to quit or fire you if you do not do as I say.[52]

However, there do seem to be cases in which it is impermissible to threaten to A, even though it is permissible to A. Consider the following:

(1') If you don't sleep with me I'll fire you from your job
(2') If you don't give me a raise I'll tell your wife about your affair
(3') If you don't give me $10,000 I'll tell the cops about your criminal acts
(4') If you don't break off your engagement I'll cut you out of my will
(5') If you don't vote for me I won't renew your contract
(6') If you don't give me $10,000 I'll join a political party you dislike

These cases seem to involve the making of impermissible threats – threats which might be regarded as extortionate, or as instances of blackmail. But we can imagine scenarios in which it is permissible to carry out the declared unilateral plan. If the employee in (1') is habitually late or incompetent, or if the terms of employment are "at will," then it seems permissible to fire him or her. If the contract is coming up for renewal in any case, and that the intervener has the moral and legal right not to renew it. If you've seen someone commit a crime, then it's permissible to report them to the police; what's not permissible is to threaten to report them to the police unless they pay you money.[53] The challenge is to explain how this could be the case.

Another way of putting the point is this: hard bargaining involves threatening to do what you have the right to do anyway; but so does blackmailing. If hard bargains should be excluded from the class of impermissible threats on the basis of the Inheritance Principle, then so should blackmail. But this is a mistake, for blackmail *does* involve making impermissible threats.[54]

But this dismissal of the Inheritance Principle might be too quick.[55] It rested on the assumption that the interveners had the right to carry out their declared unilateral plans (firing from the job,

---

[51] See Berman (2002, 53); Wertheimer (1987, 215); Haksar (1976, 71-2).

[52] It has some obvious counter-examples, but they won't be relevant here. For instance: it may be wrong to threaten another because one has promised not to, or because threatening, by itself, will have very bad consequences. Thanks to Gideon Rosen for pointing this out.

[53] Assuming that there are no further features which make the declared unilateral plan impermissible.

[54] Notice that this is not, by itself, a problem for the account of coercion given by **C**. For it is open to threat-theorists to say that **C** gives the correct account of coercion, and that the Inheritance Principle correctly excludes cases of blackmail, because the wrongfulness of blackmail is not the wrongfulness of coercion. They would then have to give an account of what is wrong with blackmail which did not in turn appeal to coercion, and many have done so (Scanlon 2008 and Wertheimer 1987 come to mind). The plausibility of the overall picture will rest in part upon whether or not we think that the wrongfulness of blackmail does in fact consist of the fact that it is coercive. As will become clear below, I lean heavily on the sorts of explanation offered by Wertheimer and Scanlon; but I generalize them in a different direction.

[55] Many thanks to Steve White for pointing this out to me, and to Sam Shpall and Alejandro Perez-Carballo for helpful discussion.

reporting to the police, and so forth). But perhaps we ought not to describe the declared unilateral plan in this way. I took it to be the 'A' element in the conditional proposal "if you [recipient] don't Φ, I [intervener] will A." And I took the permissibility of Fing to be settled by the facts as they were at the point at which the proposal was made. These might both be mistakes.

Consider first that the intervener's declared unilateral plan is not, strictly speaking, to A *simpliciter*; it is to A-if-the-recipient-doesn't-Φ.[56] And this, it might be urged, is a distinct act. The circumstances under which it is permissible to A may not be unrestricted: in particular, it may not be permissible to A when the recipient doesn't Φ.[57] This suggests the following revision of the unrestricted Inheritance Principle:

> **Restricted Inheritance.** If it is impermissible to threaten to A (if another does not Φ), then that is because it would be impermissible to A-when-the-other-does-not-Φ.

So consider, for instance, the act of firing an employee who is habitually late or incompetent. It well might be permissible to do so. Now consider firing the same employee, but *after* a threat (to fire unless they slept with you) is made and then defied. If – as is surely at least plausible – it is impermissible to fire someone under these circumstances, then the Restricted Inheritance principle explains the impermissibility of *threatening-to-fire-the-employee-unless-they-sleep-with-you.*

Does the Restricted Inheritance principle account for all such cases? Well, are there cases in which one has an unrestricted right to carry out one's declared unilateral plan, but it is nevertheless impermissible to threaten to carry them out? I think there are. Consider rights involving property in one's person or physical possessions, or in forming associations: let us suppose that, in general, it is permissible for me to shave my head, or get a tattoo, or join a political party.

Now presumably there is some story of what makes it the case that I have these rights. What I find hard to understand is why, on any such story, the right in question should somehow build in an exception for cases in which I, e.g., join a political party you abhor, but only because you refused to comply with my demand for $10,000 as a condition for not joining. The features that make it the case that I have a right to join the political party (if I have it to begin with) seem to me to persist even in the case where I do so after I have threatened you and you refuse; nor does the fact that I threaten and you refuse seem to make it the case that I *lose* these rights. This is why the rights themselves seem to be unrestricted, though of course they are not absolute.[58] Nevertheless I think I wrong you if I demand money from you in exchange for not joining a political party, or not marrying someone, or not getting a tattoo (where we assume that the threats are effective because you find the threatened outcomes abhorrent). The Restricted Inheritance Principle can't explain why this is the case.

But someone who is drawn to the basic idea behind these inheritance principles might now be tempted by a further thought. Perhaps the Restricted Inheritance Principle is not restricted enough. We may wish to include, in the description of the intervener's declared unilateral plan, a reference to the motives from which the intervener would act if the recipient failed to comply with his demands. That might lead to a principle such as the following:

> **Motivationally Restricted Inheritance.** If it is impermissible to threaten to A (if another does not Φ), then that is because it would be impermissible, if the other does not Φ, to A-for-the-reason-that-the-other-did-not-Φ.

So for instance, one may fire the habitually late employee who refuses our sexual advances for being late, but not for refusing sexual advances; a candidate's political views may be legitimate grounds for voting against them; the fact that they didn't renew a contract is not. And so on.

---

[56] Thanks to Gideon Rosen for clarification on this point.

[57] To take another example, consider the state's right to punish. Clearly this is not the unqualified right to lock-people-up, but rather the right to lock-people-up-if-they-break-the-law-and-are-convicted-after-due-process, etc.

[58] That is to say: it may be permissible to infringe them, all things considered.

How plausible is this new restriction? Our answer will depend, in part, on the relationship between intention and permissibility more generally. Should the motives from which one acts be counted as part of what it is one does when one acts, in such a way as to make a difference to the permissibility of the act itself? The issues here are subtle and difficult, and I am unsure what to say about them – I feel the pull of the idea that the permissibility of an act might depend upon the motive from which it was done, but also of the arguments raised against such a view by Judith Jarvis Thomson, Thomas Scanlon, and others.

Leaving aside these more general issues, the principle does seem to get some cases right, though it might also be open to counter-examples.[59] But I do not wish to rest to much weight on them. My main complaint with these principles comes not so much from the force of putative counter-examples, but rather because I have a hard time seeing how they constitute genuine explanations.

What they purport to explain is the wrongfulness of performing a certain speech-act, namely the utterance of a threat. What they appeal to is the wrongfulness of the act the intervener *would* perform if the threat was resisted. In this way they seek to establish a connection between what Wertheimer calls the morality of proposing and the morality of acting.[60]

But is the connection so tight? Consider that one may threaten another without intending to carry it out, and a threat may succeed in coercing another even if it is a bluff; threats may be genuine without being sincere. And one may threaten another, and succeed in coercing them, even if it would be impossible (in some quite robust sense) to carry out the threat (for instance if one uses a toy gun, or a gun that isn't loaded). It seems to me very odd that the wrongfulness of the determinate speech-act which has been performed should be explained by the fact that in some counter-factual scenario, perhaps quite remote from the actual one, it would be wrong to carry out the threat.

An additional, though related, oddity is this. Surely part of the explanation for why a threat is coercive (and therefore wrongful) involves the way in which it impinges upon the recipient's choices. Threats are coercive in part because they seem to render the option of non-compliance substantially less eligible than the option of compliance; this is part of what makes them wrongful, when they are. But the Inheritance principles explain the coerciveness of threatening without appealing to the effects of the threat on the recipient's view of her choice situation, and seem to leave no room for an appeal of this sort.

Furthermore, once one considers the recipient's point of view, it's mysterious why the question whether a threat is coercive is to be settled by the permissibility of carrying out the declared unilateral plan. For the source of the recipient's moral complaint seems rather to derive from the prospect of being *harmed* if he or she doesn't comply. Suppose I am deeply sentimentally attached to something that doesn't belong to me – the house in which I grew up, for instance. If someone, knowing this, demands $100,000 for not destroying the house, it seems to me that I have been wronged – in the distinctive way characteristic of coercion – whether or not *they* own the house (and hence have the right to destroy it).

To summarize these last two sections. I began by considering **C**, the view that coercion essentially or constitutively involves threatening another with unwelcome consequences if they do not comply with ones demands. If one takes coercion to be *pro tanto* wrongful, then hard bargains are not coercive threats, for they don't seem to be wrongful at all. But they do seem, intuitively, to be threats: so **C** requires some way of drawing a distinction between permissible threats (hard bargains) on the one hand, and (*pro tanto* wrongful) coercive threats on the other.

---

[59] One from Alejandro Perez-Carballo. I tell the person I love that I will kill myself if she does not marry me. It might indeed be permissible to kill myself-for-the-reason-that-she-didn't-marry-me; but it might nevertheless be impermissible to *tell* her that, as a way of forcing her to marry me.

[60] As he points out, one may permissibly threaten to do what it might in fact be wrongful to do in the case of non-compliance. I can tell a robber I will shoot him unless he puts down my TV, but it would be impermissible to shoot him if he did not. Nuclear *deterrence* may be permissible even if a nuclear attack is not.

One way to do this – captured by the various "Inheritance Principles" – is by an appeal to the impermissibility of the act that the intervener threatens to carry out in the face of non-compliance. Blackmail cases afford *prima facie* counterexamples to these principles, for they involve impermissibly threatening what it is in fact permissible to do. Even if the Inheritance principles can be revised to account for these cases, it remains mysterious how they should explain just why the threatenings in question are impermissible. The mystery will deepen in the following section, in which I consider the relationship between wrongfully threatening another and wrongfully coercing them.

IV

It is a commonplace that one may coercively threaten another without coercing them: for 'coercion' is a success term, and the threat may not succeed in securing another's compliance, at least in the right way. To take some standard examples: the highwayman threatens to kill me unless I hand over my money; I ignore the threat and wrestle the gun away from him. Or I know that the gun isn't loaded, but his threat moves me, out of pity rather than fear, to hand him my money. In these cases I have been wrongfully threatened, but I haven't been coerced; *a fortiori* I haven't been wrongfully coerced. The wrongfulness of an act of coercion seems to go above and beyond the wrongfulness of any act of coercive threatening.

Suppose something like **C** gives the right account of what coercion is. That is: coercing another essentially involves making a (coercive) threat to inflict a penalty for not complying with a demand. **C** provides a principle of decomposition: P coerces Q just in case (a) P makes a coercive threat to Q and (b) the success conditions are met. But we just established that the wrong of *coercing* another is not simply the wrong of making a coercive threat. Now we might ask: what does this distinct wrong of coercing another consist in?

Here is a natural thought, which I will call the "threats-as-wrongmakers-thesis." If coercing someone involves making them a coercive threat, then the wrong of coercing them involves the wrong of making them a coercive threat. What has to be explained, then, is how the fact *that the threat succeeds* turns this wrong, the wrong of making the threat, into the further, graver wrong of actually coercing them. The making of the threat, after all, is the performance of a certain speech-act. We want to know what is the connection between the fact that this act is wrongfully performed and the fact that another is wrongfully coerced. What is it about the threat, and the fact that certain other conditions are met, that is able to bring into the world an act of a wholly different type (not a speech-act, definitely), and whose wrongfulness is of a wholly different kind? It's not obvious what to say here.

Now consider another point. Strictly speaking the phrase "wrong of making them a coercive threat" contains a redundancy, for recall that we took as a condition on what it is to be a coercive threat that it be *pro tanto* wrongful. If the Inheritance principles give the right account of what makes something a coercive threat, then the wrong of coercively threatening another is in turn explained by the wrong of carrying out the declared unilateral plan in some suitably described counterfactual scenario.[61] So we might think then that the wrong of coercing some is to be ultimately explained in terms of the wrong that we would commit were we to carry out our threat in that counterfactual scenario; a wrong which we do not in fact commit when our threat is successful.[62]

This strikes me as an extremely odd result. The moral complaint that lies behind the charge "He coerced me into Φing" seems like a determinate and straightforward one. Much more straightforward, at least, than the complaint "He threatened to do something, if I didn't Φ, which it would have been wrong to do had I refused, and the prospect of which made not-Φing substantially less eligible than

[61] It's important to recall that the Inheritance principles were only one way to distinguish coercive threats from non-coercive bargain threats. There is no barrier, for one who holds **C**, to giving some other account of this distinction.
[62] Though of course a perverse or callous highwayman might shoot his victim *after* he hands over the money.

Φing; and I Φed in order to avoid this threatened consequence." The morality of coercing, and its relation to the morality of threatening, seems simpler than this.

Let me suggest an alternative picture of this relationship. Rather than characterize the wrong of coercion in terms of the wrongfulness of making a coercive threat, we should explain the wrongfulness of making a coercive threat *in terms* of the wrongfulness of successful coercion. The wrongfulness of making a coercive threat derives, not from some intrinsic characteristic such as the declared unilateral plan, but rather from the fact that it is a means towards doing something impermissible, namely coercing another; threats are *vehicles* of coercion. They are similar to attempts in this respect. Shootings are not wrong in themselves; leaving aside cases of recklessness, they are wrongful when, and insofar as, they are attempts to do something wrongful, i.e. killing another person. It would be odd to say that we explain the wrongfulness of killing someone by the fact that killings are successful shootings.

We need to be careful in describing the relationship, however. Take the case in which someone threatens me but fails to coerce me, because I defy him. Does my defiance show that his threat was a failed attempt to coerce me? No: for he may have threatened me capriciously, on a whim or because of a bet, without caring about whether I complied with his demand at all. Then it would seem odd to say that what he attempted was to coerce me; the threat wasn't part of some broader plan which had coercion as its end.

In fact, even cases of successful coercion don't seem appropriately described as successful attempts to coerce: my aim in sticking you up at gun-point is not *that you be coerced*, but rather that you hand over your money, an aim which I might well prefer to accomplish by other means. (A "pure" case of attempted coercion may be something like this: the *capo* tells me that if I want to join the mafia I must show myself capable of coercing someone into doing something. That's a case in which I care about *coercing* them; I may have no independent interest in my demand being fulfilled. The perverse *capo* might in fact get me to coerce someone into doing something I wouldn't want them to do).

So we shouldn't say that threats are wrongful because they are attempts to do something wrongful, i.e. coerce another. Rather, they are elements of a plan which, if successful, would be wrongful. The "wrong-making" features are not intrinsic to their being *threats*, but rather are a function of their role in a larger, impermissible plan. Whether they are sincere, or whether the intervener has the right to carry out the declared unilateral plan, or had the ability to do so, is irrelevant to whether or not they can play this role. For the role they play is a causal one – and the causal efficacy of a threat depends merely upon whether it is credible or not, i.e. whether the recipient *takes* it to burden their options in a way that renders non-compliance substantially less eligible than compliance.

What one might call the "threats-as-means" thesis helps illuminate a number of different issues. First, it gives us a new way to draw the distinction between hard bargains and cases of blackmail. Consider the threat, by a manager to an employee, of the form "I'll fire you unless you Φ." Compare the cases in which 'Φ' names "come to work on time" and "sleep with me." The first threat is permissible while the other is not. On my account, the explanation is given by the fact that in the first case an employee who comes to work on time in order to avoid being fired will not have been wronged, while an employee who sleeps with a manager to avoid being fired *will* have been wronged.[63]

Now consider the case of blackmail. Inheritance principles explain the wrongfulness of a coercive threat in terms of the morality of what would happen if the demand is refused, and the penalty exacted. Since the primary focus here is on the nature of the penalty, it is no surprise that cases of blackmail should seem to pose a problem, for these are cases in which the intervener seems to have a right to inflict the penalty. But the threats-as-means thesis allows for a more fine-grained description of these cases. We explain the wrongfulness of a threat in terms of the morality of what does happen when the

[63] How to characterize these wrongs is a further question, which I address in the following sections.

demand is complied, and the penalty avoided: here demand and penalty are treated on par. The wrongfulness of blackmail is explained by the *impropriety* of using the threat of penalty as a means of securing the demand in question.

The threats-as-means thesis also helps us understand cases of so-called permissible coercion. On the traditional view, the threat of physical violence is the paradigm case of a coercive threat which is *pro tanto* wrongful; threats of physical violence *never* constitute hard bargains. But now consider the threat of physical violence against a would-be rapist, or against a hostage-taker, made in order to deter the crime or have the hostages released. It seems to me that one who makes such threats (whether or not we label them "coercive") has done nothing wrong *at all*. Why? Because if the recipients of these threats comply with them, they have not been wronged. Whether or not they would be wronged were they to resist is simply not relevant to the question of whether they are wronged when they do comply. A similar point arises, I believe, in discussions of whether the law is coercive, in a way that requires justification. What requires justification is the right to punish those who disobey the law; but that is not a question about the right to *threaten* to punish. When the possibility of punishment deters me from doing some criminal act, I do not think I have been wronged in some way that calls for justification.[64]

The threats-as-means thesis also allows us to understand the distinction between threats and warnings, and why the former are impermissible while the latter are not, even though both acts involve the communication of a certain conditional intention: that they will do A if the recipient doesn't Φ. It has seemed to many that while it is impermissible to threaten, it is not impermissible to warn, whether or not the content of what is communicated is permissible or not.[65] We can formulate the distinction as follows: permissible warnings involve the communication of conditional intentions where the act of communicating is not itself part of a plan to get someone to do something. So they aren't means towards doing something that constitutes a wrong against a recipient who adjusts their behavior in the light of such a communication.

But this point can be made without appealing to any deep facts about the nature of warnings as such, and so we can make it without entering into complicated analyses of threats and warnings which appeal to differences in the intentions and aims involved. Furthermore, many threats can be converted into warnings: all one needs to do is set things up in such a way that the intervener's Aing will be triggered by the recipient's refusal to Φ, though without at that time requiring any direct action by the intervener. There doesn't seem to be a morally significant difference between the following acts: announcing the sincere intention to shoot someone if they cross the bridge (a threat); and first setting up an automatic weapon which will be triggered when anybody tries to cross the bridge, and then warning them of its existence. We assess the morality of both speech-acts in exactly the same way: i.e. in terms of the permissibility of the overall plan of which they are parts.

V

In introducing the threats-as-means thesis I have been appealing to the notion of the moral complaint someone might have when a threat succeeds. Roughly, the suggestion is that it is wrongful to threaten another when, if the recipient complied because of the threat (i.e. in order to avoid the threatened penalty), they would have a legitimate complaint against the intervener. It is time to examine the nature of this complaint more closely. I want to suggest that 'coercion' does not name a single kind of wrong, but rather a family of wrongs with the same structure. But I will begin first by considering the nature of the wrong of blackmail.

Consider Thomas Scanlon's explanation for how to explain the permissibility of a threat to fire a subordinate for not sleeping with you, when firing her as such is permissible (because, let us, assume,

[64] See Warren Quinn (1985) for an attempt to derive the right to punish from the right to threaten.
[65] As AJ Julius pointed out to me, if one is planning to do something impermissible, one might well have an obligation to warn others of this fact. He in turn credits the point to Steve White.

the employee is habitually late).[66] The idea amounts to this. There are good reasons to want some hierarchical employment structures in which people high up in the hierarchy have discretion over the careers of people lower down in the hierarchy: reasons having to do with efficiency, for instance. But this discretion must be constrained by the very reasons that make these institutional structures a worthwhile thing to have. One of these constraints might be that such discretion can be only be exercised for reasons having to do with abilities connected with the job. That is one reason why it is wrong to make an employee's keeping her job conditional on her having sex with her superior. But the structure of the justification is one that applies equally well to exclude, e.g., racial discrimination or nepotism or the case in which continued employment is made conditional on helping the superior's son with his homework.[67] Scanlon calls these "abuse of discretion" cases.

Notice that no reference is made here to the notion of a threat, or to the impermissibility of the intervener's declared unilateral plan. Nor is any reference made to how badly things will go for the recipient if she does not comply with the demand. The fact that the recipient's situation is dire is relevant to the question of how likely it is that she will comply, but it is not part of what *makes* it wrong to make the proposal, or what makes it wrong for the employer to sleep with her if she chooses to submit to it.

In fact, the point can be generalized further. For one can abuse one's discretion without making proposals at all. Giving the employee the choice: "sleep with me, or I'll fire you" is an especially egregious abuse. But the same kind of wrong would be committed by a supervisor who fired the employee without making such a proposal, but in retaliation for rejecting his sexual advances at some point in the past.[68] So one aspect of the wrong of threatening her – distinct from the threats-as-means thesis – is that it would constitute an abuse of a discretionary power, the power to threaten to fire the subordinate.

But the source of the power to threaten to fire is the same as the source of the power to fire. And in both cases, we may assume, the power isn't abused when it is exercised for reasons connected with the job itself: i.e. to make sure an employee comes to work on time, or to fire an employee who doesn't come to work on time. The Inheritance principles sought to derive the wrongfulness of the threat to fire someone from the wrongfulness of actually firing someone if they didn't comply. But in fact we should explain both in the same way, by appeals to the goals and rationale of the underlying domain. That is why the threat to fire an employee unless they help you with your child's homework is wrongful in an ordinary employment context. But surely it is not always wrongful. In particular, it is not wrongful when the person in question has been employed to help your child with their homework.

The point is a general one. There are situations in which we have power over another person, but there are constraints on how it can be exercised. What these constraints are will depend upon the particular decision we have in mind, specifically, on the kind of thing the decision is a decision *about*. For instance, we might think that it is important that the decision whether or not to vote for a particular candidate in some election should be made on the basis of assessments of the candidate's suitability for the office in question. Threats to exercise this discretion in one way or another have their place: one can threaten to report an affair unless the husband calls it off, or to vote for someone else unless this candidate changes their policy on some matter in which we have a legitimate interest.

---

[66] In Scanlon 2008, 83-86. Because he subscribes to something like the Inheritance Principle for wrongfully coercive threats, he can't say that the supervisor's proposal is wrongful because coercive; hence the need for an alternative explanation.

[67] Of course, it seems *worse* to make the job conditional on having sex rather than helping with the son's homework. See the discussion below for thoughts on why this might be so, which draws upon Scanlon's (1986) discussion of the significance of choice. A lot seems to depend upon the *value* of having control over what happens to us in various domains of our life. If one regarded sexual activity as a form of labor just like anything else, one might not have such strong intuitions about how much worse it is to ask for sexual favors than to ask for help with a child's homework.

[68] Thank to Seana Shiffrin for discussion of this point.

So the question is not whether or not we threaten, or whether the courses of action we propose if our requests are not complied with substantially burden another's options (for we may imagine in such cases that they do). What makes threats wrong is that they are abuses of discretion; what makes them abuses of discretion is that they are improper ways of getting people to do things in the underlying domain or context. And this is, as we shall see, is a substantive moral question which must be settled by reference to the aims and rationale of the underlying domain. The nature of the penalty constrains the kinds of demands one can make with a threat to exercise it – but only relative to a domain.

But this discussion is still incomplete,. For a threat to be an abuse of discretion it is enough that the demand be inappropriate, given the context in which the threat is made. At this level of description the demand for sex is on par with the demand for help with a child's homework. But the nature of the demand – the fact that it is a demand for *sex* – surely has a special salience in the moral assessment of the sexual harassment case. Now requests for sex are not in themselves impermissible, but neither are requests for help with a child's homework; nevertheless, *demands* for sex are much worse than demands for help with a child's homework. We must account for why this is the case.

The account I wish to offer draws heavily from Thomas Scanlon's discussion of the "Value of Choice."[69] Scanlon identifies three ways in which it can be important to have control over what happens to us with respect to some aspect of our lives, where the control in question is exercised by making certain choices and having them respected by others: choice may have instrumental, symbolic, and demonstrative value. The instrumental value of choice in the realm of sex is enormous: having control over when to have sex, with whom, and in what circumstances, makes a great deal of difference not only to how sex is experienced, but also to some of the physical concomitants of sex, such as pregnancy or the risk of contracting sexually transmitted diseases. The demonstrative value of choice in the case of sex is also very great: it is only because we have control over our sexual lives that we can use sex as a means of communicating love or affection, for instance. Finally, the symbolic value of having this control is very great as well. When we value control over our sexual lives because of its connection with central aspects of who we take ourselves to be, and how we express intimacy and affection, the *fact* of being able to assert such control also has tremendous symbolic value.

Two elements of the account now become crucial. First, the value of choice varies in importance across different domains; this explains the difference in salience between the demand for sex and the demand for help with a child's homework. Second, whether or not some particular act of choosing is in fact valuable in these respects will depend upon contingent features of the circumstances in which the choice is exercised. In particular, when someone exercises the choice to have sex with another because they fear the loss of a job or livelihood if they refuse, then they are having to exercise the choice for reasons which aren't among the reasons which made the choice valuable to have in the first place.

The moral complaint of the employee who has sex with a supervisor is that he has intentionally undermined the conditions under which the choice to have sex has the value associated with that kind of choice, for the purposes of having sex with her. He does so by burdening her options by means of a threat; but the wrong is not constituted by the burdening of options per se, but by the fact that reducing these options is illegitimate. The kind of demand constrains the sorts of inducements one can give another person. A threat to leave a sexless marriage may not be wrongful in this way. And there is a parallel here with the earlier discussion of warnings: one who fires an employee for being late, but in order to then offer to hire her back on the condition that she sleep with him is perpetrating the same kind of wrong as one who makes the threat.

[69] See Scanlon (1986). I discuss these issues concerning sex further in "The Significance of Consent."

What is true of the decision to have sex is also true of decisions in other parts of our lives – whom to marry, whom to leave money to, whom to vote for; but the reasons why these choices should be protected will vary from case to case.

So here is a proposal regarding the wrongfulness of blackmail: there is no "problem of blackmail" when we examine the cases at the correct level of abstraction. When we examine these cases at a more concrete level of analysis, it is easy to see what makes the intervener's proposals wrongful: the reasons it's wrong to make the reporting of crime conditional on payments from criminals will flow from a theory of our civic duties in this domain; the reasons why it is wrong to threaten to reveal another's secrets unless they vote for us will flow from a theory of why voting matters, from which will follow an account of the sorts of reasons that are an acceptable basis for voting for another person, and so forth. Once we stop treating blackmail as an independent moral category with some explanatory power, and abandon the search for some general criterion in virtue of which all acts of blackmail are wrong, we will stop being puzzled by blackmail. There's simply no general answer to the question of what's wrong with it.

VI

What then of coercion? On traditional threat-based views of coercion, it seemed as though we had a nice account of what coercion is and why it is wrong: coercion constitutively involves reducing another's options in a way that is somehow unjust, where the account of which ways of option-reduction were unjust was to come from something like the Inheritance principle. Blackmail was an embarrassment to such views, for cases of blackmail seemed in many respects like cases of coercion, but could not be captured by such a theory. And so one might want there to be two theories in this domain, one to capture the central cases of coercion, and another to capture the deviant cases of blackmail. I want to suggest that we take *blackmail* that is the central category, and then explain the wrongfulness of central cases of coercion in exactly the same way as the wrongfulness of cases of blackmail — which is to say, in different ways in different cases, and not by appeal to a single general principle.

Consider the paradigm case of wrongful coercion: the highwayman's threat of physical violence unless the recipient's hands over his money. What makes the threat wrong, I suggest, is that the highwayman who takes money in this way has stolen it. But the story of why the threat of physical violence should constitute an impermissible way of transferring property will have to appeal to the underlying rationale of the property regime itself. What means are restricted will depend in part upon the reasons why we think it important, in the first place, to exercise some control over when and to whom we transfer rights to property. It is a mistake to think that the exchange is a form of theft because it is coercive, and coercive because it involves a threat of physical violence. Rather, the threat of physical violence is a form of theft because of features having to do with the underlying justification for the regime as a whole.

One can imagine property regimes in which threats of physical violence are not always ruled out as acceptable modes of property transfer. Suppose there is in some society a ceremonial staff which by tradition must stay in the possession of the strongest adult male in a certain age range. Transfer of possession requires a challenge from another contender to engage in an elaborate – but highly regulated – kind of physical combat, and a person who refuses to engage in such combat must automatically hand over possession. Part of the challenge in these combats is to proceed by means of intimidation, i.e. to get the possessor to hand over the staff out of fear of the physical violence, without having to engage in the violence itself.

But then a threat of physical violence may be viewed (within the society) as a perfectly legitimate means of transfer of the staff in question; at least, a criticism of any particular transfer on the

grounds that it is *coercive* is neither here nor there.[70] For a less fanciful example, suppose we bet each other $50, conditional on the outcome of a boxing match between us. At the end of the third round I tell you that I will punch you really hard in the next round unless you give up just now. The threat of further physical violence makes you concede the match right away. But it does not follow that my taking away the $50 from you is thereby illegitimate.[71]

All of this suggests a picture of coercion quite different from the one with which we began. Roughly, speaking, an act is coercive (or at least, coercive-in-the-wrongmaking-sense) just in case it involves getting someone to do something by means of a conditional proposal in an improper or wrongful way. There is no doubt a role here for option-reductions: what makes some conditional proposals improper may simply be that they reduce options which an agent ought not to reduce in the circumstances and domain in question. But their role is derivative: it is mediated by an appeal to the norms governing the relevant domain. Within these domains, in turn, we appeal to the value of being able to make certain decisions or choices for certain reasons, and it is a story of what these values and reasons are which will ultimately determine which forms of influence are impermissible.

Coercion, on this view, is a sophisticated form of corruption, and coercion by means of threats of physical violence is an especially *effective* form of corruption, but only contingently so. The same sort of wrong may be perpetrated in many contexts by means of offers as well as threats. For instance, consider a prisoner offered the option of a reduction in her sentence if she agrees to participate in a dangerous medical trial; or a suspect is offered a plea-bargain, given the option of pleading guilty to a lesser charge rather than going to trial for a greater one. Debates about the legitimacy of these proposals are frequently cast in terms of whether or not they are coercive; and debates about whether they are coercive get cast in turn as debates about whether the proposals in question constitutes threats or offers. In order to make this distinction, threat-theorists appeal to an account of the appropriate baseline with respect to which we should assess the proposal. But what baseline is appropriate will, in turn, depend upon a question of what specific duties parties owe to one another with respect to what is being asked or demanded – for instance, whether the outcomes of plea-bargains satisfy requirements of justice.[72]

But if the question of whether a proposal is coercive (hence illegitimate) bottoms out into substantive, domain-specific moral inquiry it's hard to see what we gain by asking the question to begin with, and furthermore, by trying to answer it by appeal to the notion of a threat. Indeed, once we have this framework in place, there is no need to restrict the relevant modes of influence to those consisting of conditional proposals communicated by an intervener to a recipient. We should extend this to *any* way in which one person can get another person to do something – e.g. by the use of direct physical force, as well as deception, psychological manipulation, and brainwashing. I think that this is as it should be. As it happens, we often do designate certain acts of deception as coercive; if I am right, there is a perfectly straightforward and non-metaphorical reading of such claims on which they turn out to be correct.[73]

Writers in the Nozick tradition have been at pains to emphasize the distinction between 'volitional' coercion and mere physical forcings: threats proceed by affecting the recipient's options, but nevertheless leave the recipient free to make a choice between the constrained options (see for instance Pallikkathayil 2011). This is not the case with a victim whose arm is twisted so that they drop

[70] Whether or not the means of transfer is viewed as legitimate does not, of course, settle the question of whether it *is* legitimate. The point is that the focus of moral criticism will be on the *practice*, and not on individual instances of the exercise of this threat; and furthermore, that the basis of moral criticism will not the *nature* of the mode of influence in question – i.e. the threat of physical violence – but rather more general issues about what justifies the practice as a whole.
[71] In "Deception and the Structure of Moral Principles," I discuss some pertinent issues in the context of lying rather than threatening.
[72] See Wertheimer 1987, 207.
[73] See, for instance, Bok 1989, 18-22.

their bag; that person has not even *acted*. While they have been surely right to emphasize this point, given the enormous confusion in both the law and in philosophy that has arisen from ignoring it, there is a danger in the other direction as well.[74] It is no accident that people have regarded physical compulsion and coercion-by-means-of-threats as being in some sense part of the same moral category, and an account of coercion should be able to make sense of this fact.

So insofar as we care about coercive proposals because they are wrongful, and wrongfully coercive proposals are wrongful in light of domain-specific moral considerations, one might find that those domain-specific moral considerations generalize to modes of influence other than those which proceed via conditional proposals. One may relieve another of his money by cheating him as well as by threatening him with his life: both of these are instances of wrongful *theft*, and in each case the money so obtained does not properly belong to one.[75]

In fact, the law of theft has proceeded in exactly this direction. As George Fletcher (1978, 3-49) points out, the traditional law of theft regarded robbery, burglary, larceny, and embezzlement as very different sorts of crimes, because the focus of attention was on the *form* that the action took, e.g. whether or not it was 'manifest'. As Fletcher points out, the "unification" of theft law was the result of shifting focus from the form of the act to the protection of the underlying legal interest. My point is that once we recognize this interest and understand its underlying rationale, we already have the tools with which to determine whether or not a particular mode of influence is illegitimate.

VII

So we should reverse the usual order of explanation: acts are not "wrong because coercive." Rather, certain ways of getting people to do things (what I shall sometimes call "modes of influence") are correctly classified as coercive in certain contexts – namely, when it is true that, other things being equal, it is wrong to get people to do things in that way. The term 'coercion' is then defined in terms of its role in explanations of why certain acts are wrong. Different modes of influence count as coercive in a context depending upon whether they are realizers of this role.

It is clear, on such an account of coercion, why it is that, other things being equal, one who coerces another wrongs them: the conclusion follows trivially from the way in which the term 'coercion' is defined. But there is a substantive question in the vicinity, since we might ask what further features of the act ground this wrongfulness or impropriety: we might ask, that is, of any candidate coercive act, what makes it the case that *it* realizes the role in question.

But this may still seem unsatisfactory. Part of our interest in coercion is surely driven by the thought that coercive acts have certain normative properties *in virtue* of the fact that they are coercive. One reason to investigate the nature of coercion, to say what coercion *is*, is to shed light on this explanatory connection. The usefulness of 'coercion' as a term of moral and political criticism might also seem to derive from the thought that coercion claims ground moral and political claims. Thus one might object to a particular tax regime on the grounds *that it is coercive*.

If I am correct, the force of such claims must be moderated. It is perfectly legitimate to criticize a particular tax regime on the grounds that it is coercive (rather than on the grounds, say, that it will hurt the economy). But in order to establish that it *is* coercive, one will first have to settle a prior moral question. Would it be wrongful for a state to use such-and-such conditional proposals as a means of getting citizens to part with their money for such-and-such purposes? This requires independent moral

---

[74] The most egregious examples of the confusion arising from the suggestion that coercion involves something like "overpowering another's will" arise in the law of rape, which long refused to recognize coerced sex as rape precisely because it didn't fit this model. See Schulhofer (1998) for an excellent discussion.

[75] Of course there are distinctions to be made within this broader category: some impermissible ways of getting someone to do something (or of gaining possession) may be worse than others. I have no objection in principle to using a single word like *coercive* or *theft* to single out the especially bad ways. The danger is that in doing so one might obscure the fact that there is a unified explanation for why these modes of influence are improper in the context in question.

argument, which one can engage in without invoking the notion of coercion at all. One might well conclude, as a result of such argument, that the use of such conditional proposals in such a context would *not* be wrongful (for instance, because they are a necessary means of securing a just distribution of resources). If this is the answer to the underlying moral question, then the regime will not be coercive.

It would be a mistake, on my view, to think that the coerciveness of a tax regime provides *independent* grounds for moral criticism, which could somehow be 'balanced' against the grounds in favor of the regime provided by the fact that it secures a just distribution of resources (even if the final balance of reasons was such as to favor the regime, all things considered).[76] The claim that the regime is coercive only *reports* that there are moral grounds of a certain kind for criticizing it; substantive criticism of the regime would require identifying what those grounds are.

The view of coercion here defended has a parallel with "bundle theories" of the right to property. The 'bundle theory' of property holds that nothing is involved in owning something beyond having a set of more particular rights over that thing – rights to use, exclude, and transfer, for instance. But there is no *deep* fact which unifies this set of rights: they can and frequently do come apart. (We have rights to use but not transfer airplane tickets, for instance; to use but not to exclude another from one's land when they have a right of way). Two consequences follow from this way of thinking about property.

First, the question whether or not someone 'owns' something will sometimes be indeterminate, and may not be particularly interesting: the set of rights we have over corporeal and incorporeal bits of this world is not settled, in the first instance, by their status as things we *own*. Rather: whether or not we can be said to own something follows from more particular facts about which rights we have with respect to it.

Furthermore, the fact that we own something can be eliminated from an explanation of why we have some more particular right with respect to it – if to own something just is to have a set of more particular X, Y, and Z rights with respect to it, and we want to describe what particular rights we have with respect to it, then we can simply cite the X, Y and Z without further citing the fact that the bundle together constitutes the right of 'ownership'. We cannot then say, of some particular right in the bundle, that we have it with respect to something *because we own* that thing. When Rachel Maddow asked Rand Paul what he thought of civil rights legislation prohibiting the exclusion of people from hotels on the basis of race, he replied that the answer depended upon whether this constituted a restriction on the property-rights of hotel-owners.[77] That, from the point of view of the bundle theorist, is a mistake.

I take 'coercion' to be like 'property' in the following sense. It follows, on the bundle theory, that to say we have a property right in α does not tell us which particular rights we have with respect to α, on the basis of which we can correctly be taken to have a property right in it. Similarly, to call an act coercive is not yet to give the particular features of the act in virtue of which it is (other things equal) a wrongful way of getting someone to do something. All coercive acts involve improperly getting another person to do something. But this is not yet to tell us *why* the acts in question are improper ways of getting another person to do something. There may be no perfectly general answer to that question: different acts may be improper for different reasons in different cases. And so frequently the fact that an act is coercive may be eliminated in an explanation for why it was wrongful: one might simply cite the more particular reason why the act in question was an improper way to get someone to do

[76] See "Deception and the Structure of Moral Principles" for an extended (and critical) discussion of the metaphor of balancing moral reasons.

[77] See the interview of Rand Paul on the *Rachel Maddow Show*, 19 May 2010, transcript available at http://www.msnbc.msn.com/id/37252841/ns/msnbc_tv-rachel_maddow_show/. Thanks to Gideon Rosen for this example.

something in that context – where the answer *because it is coercive* is ruled out on grounds of circularity.

The force of this eliminitavist conclusion should, however, be moderated. I do not wish to deny that there are genuine "coercion-facts," or that the term 'coercion' has a useful role to play in our moral vocabulary. Nor do I wish to deny the validity of entailments, and the rationality of inferences, of the following form:

1. X is coercive
2. Absent special justification, coercive acts are wrongful
3. There is no special justification for X
4. Therefore X is wrongful

What I wish to deny is that the deductive structure of an argument of this form mirrors a relation of explanatory priority.[78] The fact that an act by which P gets Q to Φ is coercive is a consequence of the fact that it has a certain moral property (viz. that, other things being equal, it is wrong for P to get Q to Φ in that way); but it does not *ground* this property. The mistake one makes when one claims that an act is wrong 'because' coercive is like the mistake economists sometimes make (in unguarded moments) when they say that so-and-so did something 'because' it increased their utility; for utilities themselves are taken in the theory simply ways of representing an agent's underlying preferences. To say that one option has greater utility than another is simply to report that the agent prefers it to the other; it carries no independent explanatory weight.

But this does not imply eliminativism at the level of discourse: we needn't get rid of the term 'coercion' in our moral thought and talk, any more than we need to get rid of the term 'property.' The claim that an act is coercive represents that it is wrongful in a distinctive kind of way – not wrongful in the way that a promise-breaking is wrongful, but wrongful in a way that has something to do with the underlying value of choice. One might regard it on par with terms like 'Borderline Personality Disorder' in psychology. What it is to have the disorder is simply to exhibit a certain set of symptoms, and so strictly speaking one can't appeal to the *fact* that someone has the disorder in order to explain that he or she exhibits those symptoms. Nevertheless, it is not illegitimate to demarcate a class of symptoms in that way, if they tend to go together in characteristic ways, which might suggest, for instance, that they share a common etiology. The term 'coercion' earns its keep in our moral vocabulary because it is a useful mid-level moral concept which unifies many phenomena in ways that simplify moral communication; but its role is no deeper than that.[79]

References

Bok, Sissela (1989). *Lying: Moral Choice in Public and Private Life*, Pantheon Books: l978.

Fletcher, George. 1978. *Rethinking Criminal Law.* Boston: Little, Brown, and Company.

Frankfurt, Harry (1988). "Coercion and Moral Responsibility," in *The Importance of What We Care About*, New York: Cambridge University Press.

Haksar, Vinit (1976). "Coercive proposals," *Political Theory* 4(1):65–79.

Nozick, Robert (1969). "Coercion," in Sidney Morgenbesser, Patrick Suppes, and Morton White, editors, *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*. New York: St. Martin's Press.

[78] The distinction is familiar from discussions in the philosophy of science: while one can derive the height of the tower from the length of its shadow (plus some trigonometric facts), it is the former which explains the latter, rather than the other way round.

[79] Acknowledgments TBA.

Pallikkathayil, Japa (2011). “The possibility of choice: Three accounts of the problem with coercion,” *Philosophers’ Imprint* 11:16.
Pettit, Philip (2007). “Republican freedom: Three axioms, four theorems,” in C. Laborde and J. Maynor, eds., *Republicanism and Political Theory*.
Quinn, Warren (1985). "The Right to Threaten and the *Right to Punish*," Philosophy & Public Affairs 14:4, 327-73.
Scanlon, Thomas M. (1986). “The Significance of Choice,” in Sterling M. McMurrin, ed., *Tanner Lectures on Human Values*, vol 8:149-216. Salt Lake City: University of Utah Press.
Scanlon, Thomas M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge: Harvard University Press.
Schulhofer, Stephen (1998). *Unwanted Sex: The Culture of Intimidation and the Failure of Law*. Harvard University Press.
Wertheimer, Alan (1987). *Coercion*. Princeton, New Jersey: Princeton University Press.

Attila Mráz
CEU and NYU

Attila Mráz is a PhD student at the Department of Philosophy, Central European University; a Balzan Fellow at NYU's Department of Philosophy in 2011/12. He currently focuses on the interrelations of the concept, scope and site of distributive justice, on the one hand, and the epistemic and moral significance of disagreement for democratic theory, on the other. His secondary interests include meta-ethics, analytic jurisprudence, and the philosophy of art.

"Free Will, Consequential Responsibility and the Concept of Distributive Justice"

***Abstract**: Avoiding two confusions in debates on distributive justice restricts the role of metaphysical debates concerning free will in theorizing about justice. First, debates on responsibility-sensitive theories of distributive justice often fail to distinguish moral judgments of consequential responsibility from metaphysical facts of personal responsibility (Scanlon 1998, 2010). Due to this failure, I argue, Cohen's (1989) luck egalitarianism begs the central question of responsibility-sensitive theories of distributive justice as to which facts of personal responsibility should or should not translate into consequential responsibility. Further, some democratic egalitarians (Scheffler 2005) engage in the same confusion when they object to luck egalitarianism that its account of personal responsibility is metaphysically implausible: instead, they should object that it is morally irrelevant. Finally, I show that democratic egalitarians, unlike responsibility-sensitive egalitarians, are confused as to whether they think that there is a special kind of moral concern with consequential responsibility that exhausts our concern with distributive justice. If there is no such unique concern, or our concern with distributive justice cannot be reduced to such a single concern, the metaphysics of free will (i.e., personal responsibility) is even less relevant to distributive justice than suggested by my previous arguments.*

Contemporary theories of distributive justice are often regarded as relying on metaphysical assumptions to varying degrees. So-called responsibility-sensitive (hereinafter, RS) theories, including luck egalitarianism (Cohen 1989) and liberal egalitarianism (Dworkin 2001) both hold that resources should be redistributed to or from someone depending on whether they are responsible for actions resulting in their current resource-level. Luck egalitarians (e.g., Cohen 1989; Knight 2006, 2010) argue that this makes both the truth and the practical implications of the right theory of justice dependent on whether we have a free will, and hence can be responsible for anything at all. Liberal egalitarians (Dworkin 2001, 2011) hold that at least an incompatibilist libertarian free will is not necessary to assume for a responsibility-sensitive egalitarianism: the truth of some sort of compatibilism is sufficient. At the other extreme, so-called democratic egalitarians (Scheffler 2003, 2005; Anderson 1999) argue that resource-distributions should not be based on whether we are responsible for something because metaphysical libertarianism is probably false, and any other (compatibilist) theory of free will and responsibility would be too weak to justify the role of responsibility in redistribution.[80]

In this paper I argue that there are two methodological confusions in the debate concerning the relevance of the metaphysics of free will for distributive justice. I conclude that avoiding these two

[80] Democratic egalitarians also offer other arguments against RS egalitarianism—but in this paper I am only interested in the one just mentioned.

confusions (1) establishes the superiority of liberal over luck egalitarianism within RS egalitarian theories, and (2) fruitfully restructures the current debate between responsibility-sensitive and democratic versions of egalitarianism. The structure of my paper is straightforward: the first section addresses the first confusion, the second the second one; and the third section rebuts an objection.

1. The first confusion: personal vs. consequential responsibility

All RS-egalitarian views share the principle that "unequal outcomes [i.e., distributive patterns] are just if they arise from factors for which individuals can properly be held responsible, and are otherwise unjust" (Barry 2003, as cited by Scheffler 2005, p. 7). However, I want to argue that even among RS-egalitarians, the widespread agreement over this principle is illusory, and the illusion is due to a confusion in the use of the term "responsibility" Thomas Scanlon points to. In this section, I show that while liberal egalitarianism does not fall prey to this confusion, luck egalitarianism does—and this makes the latter account question-begging.

The source of the first confusion is a failure to distinguish between two senses of responsibility: personal responsibility (Scanlon 2010, p. 603), on the one hand, and consequential responsibility or substantive responsibility (Scanlon 1998, p. 248; 2010, p. 603), on the other. Judgments of personal responsibility are of a descriptive character: they concern metaphysical facts. They answer questions such as whether it is "really" or "genuinely" *S* who performed an action *a*, or whether *a* can properly be called an action at all performed by *S*, or rather it is a "mere event" happening to *S*. The traditional problem of the freedom of the will may directly bear on this question: e.g., on some accounts, an action *a* genuinely belongs to *S* only if *S* was free to *a* or *not-a*. Judgments of consequential or substantive responsibility are, in sharp contrast, moral judgments, and accordingly, are of a normative character. They answer the question whether it is permissible or required that *S* bear the burdens or benefits imposed on her by the consequences of an event or action *a*.[81] Problems concerning the freedom of the will may have a bearing on judgments of consequential responsibility, albeit only indirectly. For an account of consequential responsibility is a moral theory, and as all normative theories, it singles out certain descriptive facts are normatively relevant and others as irrelevant. They may or may not state that personal responsibility for *a* is a necessary condition for consequential responsibility for *a* (i.e., they may or may not require that an action for which someone is consequentially responsible be genuinely theirs), and even if they do, they will specify *why* a given account of personal responsibility is the normatively relevant one. The distinction between these two notions of responsibility is necessitated, in my view, by the logical thesis of Hume's Principle, revived by Dworkin (2011), which states that normative conclusions can only be drawn from a set of premises that contain a normative premise. As theories of personal responsibility are metaphysical, and *a fortiori*, descriptive theories, they cannot yield moral conclusions in themselves—we need a theory of consequential responsibility, with moral premises, to yield conclusions about, e.g., how we ought to distribute burdens. In the remainder of the paper, I will not argue for, but only assume the truth of, this principle of logic.

One dominant group of RS-egalitarian theories, namely, liberal egalitarianism (Dworkin 2001) is a theory of distributive justice which is best interpreted as providing the moral justification of attributing consequential responsibility for an act to those who are personally responsible for it.

[81] Here I diverge from Scanlon's (1998) account of substantive responsibility which only concerns actions. That is partly because I want to offer a formal characterization of what consequential or substantive responsibility means, while Scanlon offers a substantive conception or theory of consequential (substantive) responsibility. Note also that I do not use the term "attributive responsibility" at all (see Scanlon 1998, p. 248). In the framework I lay out above, attributive responsibility, which is necessary for justifying blame and related moral attitudes in Scanlon's view, is just another kind of consequential responsibility that applies for a specific set of burdens and benefits. This is a controversial interpretation of Scanlon that I cannot argue for here, nor do I need to, as my aim is not to present an exegesis of Scanlon's distinction between attributive and substantive responsibility, but rather to fruitfully exploit his overall schema in reinterpreting specific controversies about the substantive requirements and the very concept of distributive justice.

Dworkin's theory of personal responsibility falls into the class of "self-disclosure views" (Watson 1996): roughly, *S* is personally responsible for an act *a* if she identifies with the preference that serves as a reason to *a*. Why did Dworkin choose this account of personal responsibility rather than any other one? This question cannot be answered without considering his account of consequential responsibility. One of the pivotal moral premises in his argument for this account is that each person should, and should be allowed to, to lead a life according to her own conception of what a successful life consists in. So, when a Dworkinian liberal egalitarian assesses whether *S* is consequentially responsible for an action, she is interested—among other things—in whether the action reflects *S*'s values and conception of what a successful life is like. *That* is the notion of "genuineness" that a liberal egalitarian account of personal responsibility needs to establish, and if personal responsibility along these lines is established (i.e., it is the case that *a* genuinely belongs to *S*), one—but only one—necessary condition of consequential responsibility-attribution is satisfied.

Liberal egalitarianism, in my interpretation, offers a positive argument for an equal distribution of resources because it caters for the *moral* intuition that even if our actions are genuinely our own in the required sense—we identify with them—it is unfair to hold us consequentially responsible for them if we weren't equally well positioned when we had to make decisions about what kind of lives we will lead. The argument for this unfairness-claim is complex, and I do not intend to investigate it in any detail here. Suffice it to say that liberal egalitarians offer an argument both for the egalitarian aspects of the distribution of resources and for deviations from material equality, by providing a complex moral theory of consequential responsibility which establishes the moral relevance of a specific theory of personal responsibility as one necessary condition, and fairness in terms of initial resource equality as another, for consequential responsibility-attributions.[82]

Another dominant group of RS-versions of egalitarianism, namely luck egalitarianism (e.g., Cohen 1989) also seems to assume that a theory of justice is identical to a theory of consequential responsibility. Yet luck egalitarians fail to make a distinction between personal and consequential responsibility. They merely assume that establishing the metaphysical, descriptive fact of some sort of—rather demanding, incompatibilist libertarian (e.g., Cohen 1989, pp. 927–934) or at least an exclusively backward-looking compatibilist (Knight 2010)—personal responsibility for an act is sufficient for establishing the moral, normative fact of consequential responsibility for that same act. This assumption is warranted within the theory only by an equivocation on the two uses of the term "responsibility".

One peculiar consequence of this strategy is that, as Susan Hurley (2003, Ch. 6) observes, luck egalitarians unfortunately fail to establish why we should distribute anything *equally* in any circumstances at all—one might add, except if hard determinism is true, and hence a metaphysically libertarian account of personal responsibility some luck egalitarians hold yields the conclusion that no-one is ever responsible—personally, and according to the luck egalitarian, at once consequentially—for any action at all.

However, while granting Hurley's worry that luck egalitarians are rather luckeans than egalitarians, I want to go further. Not only do luck egalitarians fail to argue for an egalitarian distribution, I claim that they also fail to argue for *any* distribution of any kind.[83] Their account begs every important question in distributive justice conceived as a theory of consequential responsibility, as they merely assert that some fairly specific account of personal responsibility is necessary and sufficient for consequential responsibility. But as Hume's Principle requires, we must argue for such propositions by means of

[82] There are, or at least, depending on your interpretation, there may be further necessary conditions of consequential responsibility on Dworkin's theory—again, my aim is not to provide a detailed interpretation, but to call attention to an important feature of the general structure of liberal egalitarian theories of justice.

[83] I am concentrating on the Cohenian version of the view, which exhibits the characteristics I am interested in in their most acute forms—some other versions may be vulnerable to the following critique to varying degrees, but I do assume that all extant forms of luck egalitarianism *are* vulnerable to it to some degree.

providing a *moral* argument; for instance, by showing that it is fair to hold people consequentially responsible for the acts for which they are personally responsible, or that fairness is irrelevant to consequential responsibility (which would be a rather implausible, yet also a substantive moral claim itself). Of course, arguments to that effect could be offered, and thus the account could be strengthened.[84]

I conclude that liberal egalitarianism is a superior form of RS-egalitarianism in comparison to luck egalitarianism, for the former, but not the latter, provides moral arguments for a theory of consequential responsibility—a moral theory—, and offers a theory of personal responsibility which spells out "genuine" action in the sense of genuineness required by those moral arguments. Luck egalitarianism puts the cart before the horse by first providing a metaphysical theory of personal responsibility without clear moral motivation for its relevance to any theory of consequential responsibility, and then merely asserting that personal responsibility is necessary and sufficient for consequential responsibility.

2. The second confusion: consequential responsibility vs. distributive justice

In the remainder of this paper, I want to argue that there is another unclarity pervading debates between theories of distributive justice. This unclarity concerns the relation between theories of consequential responsibility conceived of as a special moral concern, on the one hand, and theories of distributive justice, on the other. Spelling out this relation has a crucial methodological consequence concerning the relevance of the free will debate for theories of distributive justice.

RS (luck- and liberal) forms of egalitarianism offer an implicit reductive analysis of the concept of distributive justice in terms of consequential responsibility conceived of as a distinct kind of moral concern. That is, the right account of distributive justice just *is* (identical to) the right account of a unique moral concern which bears on when people are consequentially responsible for their acts. Dworkin's liberal egalitarianism as a broader political theory holds that for ethical and moral reasons, people ought to be consequentially responsible for the acts they identify with, and then spells out in its account of distributive justice the conditions in which this can be fairly done. Cohen's luck egalitarianism, in turn, is more mysterious, as I have argued: it simply asserts that people should be consequentially responsible for the choices properly ascribed to them on some specific metaphysical account of choice and free will.

In contrast to RS forms of egalitarianism, however, democratic egalitarianism is less clear about whether or not it assumes that a theory of distributive justice is exhausted by an account of a distinct moral kind which bears on judgments of consequential responsibility. If it shares this assumption, it is best interpreted as a rival moral account of consequential responsibility competing with RS egalitarianism on the same level of analysis. For instance, democratic egalitarians may insist that there is a single, distinct moral concern with political equality, and distributive justice is identical to consequential responsibility as governed by this ideal of political equality alone. On this interpretation, democratic egalitarians may object to RS-egalitarians that the moral theory of consequential responsibility the latter endorse is wrong for normative reasons, and hence it also attributes moral relevance to metaphysical facts (in particular, those related to personal responsibility

[84] Knight (2006) at least takes the right direction when he writes that "luck-egalitarianism looks far more plausible if it rewards and penalizes in a manner that is genuinely (i.e. according to the correct metaphysical theory) responsibility-sensitive[,… which assumption] is metaphysical in character but morally motivated" (p. 177), but I disagree with his claim that such assumptions are metaphysical in the first place. In my view, Knight fails to draw the conclusions from the argumentative role of the "moral motivation" in theories of consequential responsibility. Moreover, he doesn't realize that "genuinely" (personally) responsible acts are genuinely so not because they are so according to the right metaphysical theory, but because they are so according to a metaphysical theory which properly reflects the sense of genuineness relevant for a given account of consequential responsibility. See my further comments on this issue in reply to the first objection I address in Section 3.

and free will) that are morally irrelevant. For instance, democratic egalitarians might object to luck egalitarians that people should not bear all the costs of the satisfaction of the preferences they identify with because it is morally objectionable to allow people to enslave themselves even if they wish to do so and their wish is not a mere wanton's craving, nor an addict's compulsive desire; and this is so even if their choice to enslave themselves was a metaphysically libertarian choice.

Yet if this is so, we must draw an important methodological conclusion regarding the relevance of the free will debate to the RS-egalitarian vs. democratic egalitarian controversy, based on the results of the previous section. Democratic egalitarians cannot—and need not—object to RS-egalitarians that the latter rely on an implausible metaphysics of free will and choice. For instead, their objection, correctly formulated, states that the metaphysical assumptions of RS- egalitarianism are *morally irrelevant*, whether they are plausible or not. It is therefore a mistake to think that metaphysical considerations can directly contribute to the debate between responsibility-sensitive and democratic forms of egalitarianism—although this mistake is committed by some luck egalitarians (Knight 2006) and democratic egalitarians (Scheffler 2005) alike. Contrary to appearance, this methodological restriction is as true of the debate between different forms of RS-egalitarianism as it is of the debate between democratic and RS egalitarianism in general.

On the other hand, if, contra RS egalitarians, democratic egalitarians assume that a theory of distributive justice as consequential responsibility cannot be exhausted by an account of a single moral concern, then the debate between RS and democratic egalitarians cannot even be restricted to the question of the right account of that very concern. First, what RS egalitarians argue for is, on their concept, an *all things considered* theory of distributive justice, whereas on the democratic egalitarian's concept of distributive justice, it is merely—if at all—a *pro tanto* theory of justice in distribution.[85] Thus it is misleading to set up these two sorts of egalitarianism as rivals, for they are promoting these theories at different levels of analysis. Second, then, on this interpretation of the debate between democratic and RS egalitarians, their controversy must also focus, in addition to the right account of a distinct moral concern with consequential responsibility, on the very reasons for or against equating distributive justice with this and only this moral concern (if there is such a concern at all), since even if democratic and RS egalitarians settled on the existence as well as the right view of the unique moral concern with consequential responsibility, that could not in itself dissolve their controversy.

If the second interpretation of the debate is the right one, that only goes to show that the metaphysics of free will has an even more compartmentalized role in the RS vs. democratic egalitarian debate than what the first interpretation suggested—namely, it only has a potential role in justifying pro tanto attributions of consequential responsibility based on a special moral concern. Justifying consequential responsibility on the basis of such a concern, in turn, may exhaust a theory of distributive justice, if some form of RS egalitarianism is true, or it may not, if some form of democratic egalitarianism is true.[86] However, whether the justification of consequential responsibility based on a

[85] The caveat "if at all" is important, as some versions of democratic egalitarianism may well deny that there is any special, unique *pro tanto* concern with consequential responsibility. Cohenian luck egalitarians certainly think there is such a concern—roughly identical to some sort of prejusticial desert—, whereas Dworkin's strategy seems to be the provision of a reductive analysis of at least some of the other democratic egalitarian core concerns (e.g., political equality) in terms of the single concern which underlies his account of economic equality. So, Dworkin grants the apparent plurality of the moral considerations bearing on distributive justice as consequential responsibility, but he also tries to reduce most of them to a single fundamental liberal concern. In contrast, Scheffler for instance may be read as denying any kind of unique *pro tanto* concern with consequential responsibility that should be weighed against other concerns in *all things considered* judgment of justice in distribution, but he or other democratic egalitarians may be read instead as arguing that although there is such a concern, it is only one of the concerns we care about under the label of distributive justice. Both readings are coherent with my second interpretation of the democratic egalitarian position—in this paper, my main focus is on the second reading, but this expository focus should not bear on my arguments.

[86] These two types of theory may well not exhaust the logical space for possible theories of distributive justice, but as they pretty much exhaust the current related controversies in political theory, I restrict my observations to these two.

special moral concern can exhaust a theory of distributive justice is, again, a moral question which cannot be answered by further metaphysical inquiry into the nature of free will and choice.

3. An objection and replies

Wrong Theories of Personal Responsibility. You may object to my claim that free will debates cannot settle debates between different theories of consequential responsibility (and hence, distributive justice). You might concede that the structure of these theories corresponds to what I sketched above: a moral theory of consequential responsibility conjoined with a metaphysical theory of personal responsibility. But what if, you may ask, a given moral theory of consequential responsibility relies on a *wrong* theory of personal responsibility? Then the given account of consequential responsibility—and theory of justice— is false, too. Therefore, metaphysical considerations, and those concerning the right account of free will in particular, do have a decisive role in arguing for or against (at least some, namely RS) theories of distributive justice. For example, liberal egalitarianism, as I said above, relies on a real self account of the freedom of the will. If that is a wrong account, liberal egalitarianism fails.

Yet the objection is unclear. To clarify it, there are three questions we might want to know about an account of personal responsibility: (1) whether it is morally relevant, (2) whether the conditions it formulates are ever satisfied in the actual world, and (3) whether it is coherent. "Wrongness" may be interpreted along any of these dimensions. Let us consider them one by one.

First, a theory of personal responsibility is morally relevant if and only if its attribution is necessary for attributing consequential responsibility, on a given account of the latter. If a theory of consequential responsibility is morally dubious, then the personal responsibility-attributions it requires are irrelevant—but that has nothing to do with the question whether the metaphysical theory of personal responsibility that yields them is right or wrong as a metaphysical theory. It may carve up the world at its joints, point to real distinctions etc. For instance, it might provide us with the true description of what identification really is in human moral psychology. Yet, if identification is not what we care about, it is morally irrelevant, and even if an account of what identification is and when it obtains is true, this in itself can't give any reason for accepting a theory of consequential responsibility (distributive justice). But the "moral irrelevance" of an account of personal responsibility can't be the sense of wrongness the objection is after, as it is not independent from the rightness or wrongness of a theory of consequential responsibility.

Still, as for the second question, what if a theory of personal responsibility on which a theory of consequential responsibility relies singles out a feature of agency that could exist, but actually does not? Perhaps that is the issue the objection really addresses, as in this case, the theory of personal responsibility at hand is descriptively false: nothing corresponds in the actual world to the description it provides. Yet that does not undermine a theory of consequential responsibility—rather, that only implies that in the actual world no-one can be held consequentially responsible (on the given theory) for anything on the basis of their personal responsibility, as Knight (2010) correctly observes. Since theories of consequential responsibility usually have a conditional form—*if S* is personally responsible for *a*, and other conditions hold, then *S* should bear the burdens and/or benefits incurred by *a*—, the imagined case only shows that the antecedent is never true in the actual world. This, I assume, must trigger an "or else" clause of a full theory of distributive justice as consequential responsibility. For a full theory of distributive justice formulates prescriptions for all possible states of affairs.[87] If as a

[87] Two caveats are in order. First, an unrestricted universalization may seem too strong here. Available theories of distributive justice limit themselves to a rather restricted set of possible states of affairs. This is perfectly understandable given that they want to provide prescriptions for ordinary human politics, not some distant possible world. Yet, as the well-known Nozickean thought-experiment about "Utilitarianism for Animals, Kantianism for People" shows, even taking stock of our moral intuitions about possible beings like aliens may change or at least reveal some implicit premises in moral thinking (Nozick 1974, pp. 35–42). Second, if some form of moral naturalism is true, then moral facts can vary across (relevantly) different possible worlds. So, on moral naturalism, a theory of distributive justice can only be true for the set of

matter of metaphysical truth, some of these conditions never obtain in the actual world, that at most renders some parts of this moral theory practically uninteresting, but not wrong.

Third and finally, my objector might want to claim that if a theory of personal responsibility is incoherent, that would certainly undermine the theory of consequential responsibility relying on it. But that need not be so.

On the one hand, for instance, imagine that a theory of personal responsibility holds that an act *a* is genuinely attributable to *S* only if *S* identifies with it, but it conceives of *S*'s moral psychology as including no second-order preferences, nor an independent extrainclinational will. Such a metaphysical theory of personal responsibility cannot spell out what is meant by identification. Its incoherence implies that it fails to capture a notion—identification—that, we assume, is morally relevant for some theory of consequential responsibility. Yet I believe that in the given example, this failure does not undermine the given theory of consequential responsibility. As a moral theory, that only relies on the assumption that the morally relevant concept of "identification" can be spelled out in a coherent theory of personal responsibility. So, in our example, the theory of consequential responsibility that refers to identification may turn down one particular, unfortunately incoherent theory of identification, and instead rely on another theory of personal responsibility which provides an improved metaphysical understanding of this notion. So, wrong theories of personal responsibility, in the sense of descriptive incoherence, often do not undermine theories of consequential responsibility as moral theories—though the latter may call out for theoretical improvement in theories of personal responsibility.

On the other hand, there might be cases where the morally relevant feature of agency that a theory of consequential responsibility refers to cannot be spelled out in *any* coherent way in a metaphysical theory. (I suspect some may hold "agent-causation" to be such a notion, while others would go for "the self-originating will"—pick your own favorite example.) In these cases, a theory of consequential responsibility provides prescriptions which are conditional on something that cannot exist (in any possible world). Still, I argue that the upshot of such cases does not crucially differ from my conclusion about those cases in which the metaphysical condition is never met only in the actual world. Even if a theory of consequential responsibility uses an incoherent account of personal responsibility, that does not undermine its truth as a moral theory, but only guarantees the falsity of the antecedent of its conditional moral premises. In other words, the incoherence of the relevant notion of personal responsibility just renders a theory of consequential responsibility rather uninteresting. Yet we have every reason to hope that any reasonably well-developed theory of consequential responsibility does not formulate prescriptions only for metaphysical impossible scenarios, but also for possible and even actual ones.

To sum up, if the metaphysical account of personal responsibility required by the moral account of consequential responsibility is morally relevant (because we are talking about the *right* theory of consequential responsibility), coherent, and refers to an actual feature of human agency, there is no further question to ask as to whether it is *the right* theory of personal responsibility or free will.

4. Conclusion

In this paper I have argued that weeding out two confusions from debates on distributive justice—namely, one between personal and consequential responsibility, and another one concerning the relation of a special moral concern bearing on consequential responsibility, on the one hand, and distributive justice, on the other—sets limits to the role of metaphysical debates concerning the nature of free will in theorizing about justice. In particular, I found that Cohenian luck egalitarianism is guilty of overstepping one such limit by asserting without moral argument that a particular account of

---

descriptive phenomena that "ground" (in some adequate sense) those principles. I believe I must concede this point, as the framework I apply here (with the Dworkinian interpretation of Hume's Principle) has non-naturalistic meta-ethical commitments. For lack of space, I cannot here argue for this conviction.

personal responsibility is sufficient for consequential responsibility. RS and democratic egalitarians commit either the same kind of mistake when they try to undermine each others' theory by reference to their respective metaphysical commitments, or (on another interpretation of their debate) they commit an even more serious trespass when they fail to consider that part of their controversy is not even about a morally unique concern of the RS egalitarian kind bearing on consequential responsibility, and hence the metaphysics of personal responsibility is utterly irrelevant to that part of it. Of course, all of this does not suggest that the metaphysics of free will is irrelevant to distributive justice—that would be a substantive normative claim I didn't even begin to argue for. I have provided reasons, instead, for the methodological conclusion that free will debates have a much more limited role in debates over the right theory of distributive justice than it is commonly thought to be the case.

References

Anderson, Elizabeth S. (1999). What is the Point of Equality? *Ethics* 109 (2): 287–337.

Cohen, Gerald. (1989). On the Currency of Egalitarian Justice. *Ethics* 99 (4): 906–944.

Dworkin, Ronald. (2001). *Sovereign Virtue: The Theory and Practice of Equality*. Cambridge, MA—London, UK: The Belknap Press of Harvard University Press.

Dworkin, Ronald. (2011). *Justice for Hedgehogs*. Cambridge, MA—London, UK: The Belknap Press of Harvard University Press.

Hurley, Susan. (2003). *Justice, Luck and Knowledge*. Cambridge, MA: Harvard University Press.

Knight, Carl. (2006). The Metaphysical Case for Luck Egalitarianism. *Social Theory and Practice* 32 (2): 173–189.

Knight, Carl. (2010). Justice and the Grey Box of Responsibility. *Theoria* 57 (124): 86–112.

Nozick, Robert. (1974). *Anarchy, State and Utopia*. New York: Basic Books.

Scanlon, Thomas. (1998). *What We Owe to Each Other.* Cambridge, MA—London, UK: The Belknap Press of Harvard University Press.

Scanlon, Thomas. (2010). Varieties of Responsibility. *Boston University Law Review* 90 (2): 603–610.

Scheffler, Samuel. (2003). What Is Egalitarianism? *Philosophy and Public Affairs* 31 (1): 5–39.

Scheffler, Samuel. (2005). Choice, Circumstance, and the Value of Equality. *Politics, Philosophy and Economics* 4 (1): 5–28.

Watson, Gary. (1996). Two Faces of Responsibility. *Philosophical Topics* 24 (2): 227–248.

Kristina Gehrman
Miami University of Ohio

Kristina Gehrman (kgehrman@gmail.com) is currently an Assistant Professor of Philosophy at Miami University of Ohio. She received her Ph.D. from UCLA in 2011. Her primary research interests are in ethics, the metaphysics of value, and action theory.

"Action as Interaction"

***Abstract**: In "The problem of action", Harry Frankfurt cautions against what he describes as the 'parochial bias' of anthropocentrism in the theory of action. This paper proposes a characterization of agency and action that is suitably non-anthropocentric, building on Frankfurt's claim that actions occur under the 'guidance' of an agent. I claim that action is best understood as a particular kind of interaction, or relation, between an agent and the world. Correcting for anthropocentrism reveals the interactive quality of action, and allows us to better understand the role of agency in agents' lives. Recognizing that action is a kind of interaction also suggests a fresh way of thinking about the norms of action: among other things, it suggests that the world is the measure of success for action.*

In "The Problem of Action", Harry Frankfurt cautions against unwarranted anthropocentrism in the theory of action, describing it as a "parochial bias", and remarking that "we [humans] are far from being unique either in the purposiveness of our behavior or in its intentionality" (78). Frankfurt says that *the* problem of action theory is to "explicate the contrast between what an agent does and what merely happens to him" (69). But this contrast between 'what an agent does' and 'what merely happens' can be plainly seen in the lives of other animals: just as a human being can raise her hand or have her hand raised by a strong gust of wind, so a dog can wag its (own) tail, or have its tail moved by some external force. The subject matter of action theory – namely *action* – is thus not a distinctively human phenomenon.

Now, if beings besides humans are capable of action, then, Frankfurt points out, the difference between actions and 'mere happenings' (other events) cannot be explained "in terms of any of the distinctive higher faculties" (78) of human beings. Suppose, for example, that humans are rational, whereas dogs are not. Then the difference between mere happenings and actions, which occurs in both dogs and humans, cannot be explained by appeal to *reason*. The dog's tail-wagging is an action, but it is not explained by the dog's rationality since, by hypothesis, the dog lacks reason. Therefore, in order to explain what action is, we need to understand what it is that things like wagging one's tail and raising one's arm have in common, that makes them both *actions* as opposed to other sorts of occurrences.

My aim here is to present a picture of action that I believe avoids the 'parochial bias' of anthropocentrism, distinguishing actions from other events without relying on any distinctively human faculties to draw that distinction. The account I'll propose explains action as a certain kind of *interaction*, or relation, between an 'agent' (a being capable of action) and her environment. It can be tempting to think of action like a vector (or even a missile), with a point of origin in the agent, and a terminus in the world which the agent aims to impact. But I think action is better seen as an ongoing, dynamic *exchange* between agent and environment. To act is really to participate in a delicate, continually evolving coexistence. When we adopt a less anthropocentric stance towards the phenomenon of action, we can more easily perceive its interactive character. And we can thus do

justice to certain features of the phenomenon which may otherwise remain obscure to us.[88] I will talk about one of these features today: avoiding anthropocentrism about action discloses an attractive way to think about the *norms* of action.

I.

Let's begin with Frankfurt's own proposed solution to "the" problem of action. He suggests that actions are distinguished from other kinds of events by the fact that "the movements as they occur are *under the person's guidance*" (72, my emphasis). Since the notion of personhood normally only applies to human beings, I instead use the unlovely word 'agent'. A dog, a spider, and a human being can all be the *agents* of their own actions. Frankfurt illustrates this distinction with the example of a spider:

> consider the difference between what goes on when a spider moves its legs in making its way along the ground, and what goes on when its legs move in similar patterns and with similar effect because they are manipulated by a boy who has managed to tie strings to them. In the first case the movements are not simply purposive, as the spider's digestive processes doubtless are. They are also attributable to the spider, who makes them. In the second case the movements occur but they are not made by the spider, to whom they merely happen. (78)

The spider's movements *accomplish* something in its life – getting it from point A to point B – just as its digestive processes accomplish something in its life. But unlike digesting, *walking* is "attributable to the spider".

I am going to assume that *being done by an agent as such* or *taking place under the guidance of an agent* is constitutive of something's being an action. If actions are events (or behaviors) that are 'guided' by agents *qua* agents, then if we can shed light on what *agency* is, we'll shed some light on what action is in the process. So in the passages to follow, I'm going to sketch out a characterization of agency.

Now, when I say I'm going to give a *characterization* of agency, I have in mind the following. Suppose someone asks me what a pen is. I might tell them that a pen is an ink-filled writing implement of roughly the right shape and size to be held and manipulated by the human hand. In giving this explanation, I am telling the person the defining or typifying traits of pens. Or, suppose you want to teach a child what *planets* are. Part of a characterization of planets is going to involve reference to their characteristic activity of orbiting around star. In giving a characterization of agency, I will be assuming that agency, like pens and planets, is one kind of thing in the world, which is capable in principle of being better understood through careful observation and examination. I will be attempting to identify at least some of agency's characteristic, defining or typifying traits.

Let's begin by noting a broad distinction between static and active characteristics. Many characteristics of things in the world are *static* traits, like color, shape, size, or material composition. But there are also many characteristic *activities*, in a broad sense which includes many kinds of 'mere happenings', as well as actions. Being filled with ink is a static trait, whereas orbiting is a characteristic activity. Agency seems to be active, not static. So we might say that agency is a capacity for engaging in a certain sort of activity.

Now we can make a further distinction within active traits, between those which are peculiar to *living* things, and those which are not. Erupting and orbiting, characteristic activities of volcanoes and planets respectively, are not activities of living things, whereas ruminating is an activity that is

[88] As you might expect, anthropocentrism in one's conception of action impairs one's ability to explain, respect, and plan for the agency of other animals. Think for example of the various environmental disasters humans have caused by forgetting that other animals have agency: the introduction of invasive lampreys into Lake Champlain, for instance, or the damming of rivers used by spawning salmon.

characteristic of living things – goats, cows, and other ruminants. *Agency* appears, like rumination, to be an activity the capacity for which is characteristic of some living things but not others. Only ruminants ruminate, and only agents act. And not all activities engaged in by agents are activities of agency, just as not all activities engaged in by ruminants are rumination. (You and I are agents, and yet our digestive processes are not exercises of our agency.) Agency is one capacity among others, possessed by some living things among others, for engaging in a certain sort of activity.

Now, just what might this 'certain sort of activity' be? It is, we have already decided, a sort of *guidance*. What makes an exercise of agency an activity that is distinctively one of guiding? One *wants* to say that the crucial difference is the involvement of the mind of the creature in the case of an exercise of agency. It is the being's mind that is *doing* the guiding, when an event takes place 'under the guidance' of an agent. A planet has no mind with which to guide its own trajectory through space, and a plant has no (discernable) mind with which to guide its re-orientation towards a moving source of light. Therefore, planets and plants are not agents, and do not ever act, although they do engage in other sorts of characteristic activities. And, though we have minds, our minds do not control e.g. the movements of our intestines, therefore, those movements *in us* are not actions.

The trouble is that we (or I, anyway) really have very little idea what *mind* is, how it arises, and where and why it occurs. So an appeal to the mind of the agent doesn't shed all that much light on the subject. Certainly, the capacity for conscious experience *appears* to be restricted to living things. In the case of the planet, we tend to think consciousness is not part of the activity of *orbiting* for two reasons: one) we have no evidence that the planet is aware, and two) we think we can explain everything it does without appealing to consciousness. But it's not so clear that both of these things are true about say, a sunflower that turns daily to face the sun. Self-movement in response to stimuli is some evidence that there may be awareness on the part of an organism. And the sunflower moves because it's *alive*, or it moves *qua* living thing, not because it's a machine or because it's being acted on by an external force. Is the presence of a root structure, the absence of a central nervous system, and the lack of language-use conclusive grounds for ruling out an attribution of some sort of mind to a sunflower?

I don't know (seriously). The sunflower is a living thing that can move *itself*, or move *on its own*, as opposed to being moved by a farmer or a gust of wind. But perhaps there is a way of drawing a somewhat tighter circle around the activities that are truly characteristic of agency. In *Origins of Objectivity*, Tyler Burge suggests that there's a line to be drawn between creatures moving in response to stimuli (an exercise of agency), and creatures simply moving in reaction to stimuli (not an exercise of agency). The sunflower appears to be a one-trick pony, and hence to fall into the latter, non-agential category. That is, it seems unable to control (guide) its phototropic movements. If you put a (let us say transparent) obstacle in its path, the sunflower cannot back up and try to navigate around the obstacle. In contrast, a tick who moves itself towards a heat source really can be seen to guide its own movements. For if you poke the tick, it will back away, even if this brings it farther from the heat source. If you put a pebble in front of it, it will try to go around.[89] It appears to move *in response to* stimuli, not just in a way that is caused by stimuli.

Now whether we talk about it in terms of guidance, or whether we talk about it in terms of the capacity to respond and not just react, I am strongly inclined to say that *some* sort of mental life or capacity for awareness is basic to agency. The notion of guided movements implies the presence of a guide. The mind is what guides the movements when an action takes place.

But the thought that action presupposes mind may seem less controversial than it is. There is clearly a capacity for action on the part of beings who we don't usually think of as having minds or mental lives. (Ticks for instance.) If there is mind wherever there is non-reactive (guided or controlled) action, then we might have grounds for attributing mindfulness not only to ticks, but to beings as unlike us as ants, sea scallops, and possibly even some single-celled organisms. All of these creatures

[89] Burge, *Origins of Objectivity*.

have the characteristic marks of being agents: they are living things capable of guiding their own movements in a way that seems to *respond*, not just *react*, to the world they inhabit.

This is not meant to be a complete characterization of agency. But if we do attribute agency to creatures like ticks and slugs and spiders, we should note three implications. One, human exceptionalism about mind and action is absurd. Two, mind is a very common feature of the world; if it is mystifying to us it is anyway not that special. And three, *action* may be more fundamentally associated with consciousness than *belief*.[90]

II.

The point of so briefly raising these difficult questions is to emphasize that one can't *begin* to appreciate the difficulty of the topic of *practical life*, nor can one begin to see the contours of a satisfying treatment of the phenomenon, until one begins to appreciate just how wide, and varied, and pervasive, the phenomena of minds and of agency really are. Avoiding anthropocentrism about action opens up a whole range of interesting philosophical questions and puzzles that otherwise remains obscure.

And yet I think we can still make further progress towards characterizing agency, even while acknowledging our bafflement about *why* agency arises in the world and how to tell *where* it occurs. If we bring together the various notions of mindful guidance and response vs. reaction that we have just been discussing, I think so far we have something like the following characterization of agency. Agency is one capacity of living things among others, possessed by some kinds of living things and not others. It is a capacity for engaging in a certain sort of guided activity, where the being's mind is doing the guiding, as evinced by signs that the behavior is *responsive* and not merely reactive (controlled, and neither random nor invariant). We may not know why and how awareness arises, or which things are aware, or what variety of conscious experience is possible, or what it takes for a being to be able to initiate interactions. But, if a living entity *does* engage in the relevant sort of interaction with the world, then it's got agency.

I wish to emphasize that awareness or mind is not merely co-present with action. Awareness allows an agent to tell the difference between different sorts of things it confronts. It therefore *shapes* the interactions which a living thing initiates in an exercise of agency. The interactions of agents with the world are characterized by conscious discrimination, or differentiation. We might say that exercises of agency have grounds, or bases. Things *seem* to an agent to be a certain way, and in the normal case its actions are shaped by its assessments of those seemings. Thus, although action is organism-initiated, the organism draws its grounds from the world with which it interacts.

At this stage of characterization, our very *vocabulary* seems stubbornly biased towards the human. How can a tick be said to have 'grounds' for its actions, as if it could think to itself, 'uh-oh, forefinger in the way, best move over here'? Isn't having *'grounds'* for action a conceptual ability that is unique to reasoning creatures? If we stick to the initial prejudice which Frankfurt sought to undermine, we will be inclined to insist that the *only* way to act based on a discriminating assessment of one's surroundings (the only way to have 'grounds') is to use reason, concepts, or other distinctively human abilities. But if instead we stick to the phenomena, we can see plainly that the ability to *pick* what to 'go for' is not unique to human beings; even grubby little ticks seem to do it quite well. If ticks are too borderline, then certainly discrimination and differentiation are clear in the case of a cat or a dog who has a favorite sleeping place (or a favorite person). This is not to say the insect-agent (or the canine- or feline-agent) has *exactly the same* capacity to discriminate that a human-agent has. The 'picking' or 'discriminating' or 'having of grounds' that is characteristic of agency exists on an

[90] Note to the reader: unfortunately I do not expand upon the third claim in this version of the paper, and will therefore probably only mention the first two points when presenting.

uncharted spectrum of robustness, across different types and degrees of agency, and we must expect variation and gray areas.

If we can find a way to be comfortable with the idea that exercises of agency characteristically have grounds, or are shaped by discrimination and differentiation, then we can add the final element of our characterization of agency. Frankfurt pointed out in the passage noted above (78) that a spider's digestive processes are, as he calls it, *purposive* in the sense that they accomplish something in the life of the spider. Or, we might think of vision. Vision gives an organism with the capacity for sight a visual representation of the world; seeing is a mechanism for getting a certain kind of information and/or having a certain type of experience. Presumably agency too has a characteristic point or purpose in the lives of beings who possess it, just as rumination or digestion or sight have their characteristic point or purpose.

I propose that the basic purpose of agency in the lives of beings who possess it is that *things go better,* and that they go better precisely by dint of the direct, deliberate, and discrimination-based *influence* that agency permits the creature to have over its own circumstances. When a being interacts with its environment in the ways that are characteristic of agency, the being's conscious participation in the flow of events takes place in accordance with how things in the environment strike it. To exercise agency is therefore (characteristically, not invariably) to interact with the world in a way that one somehow takes to be warranted, fitting, worth doing, or called for – based on one's appreciation of the situation.

As with the idea that actions have grounds, the idea that the purpose of action is that things go better cannot be construed in human-specific terms. A tick, we may safely assume, does not in *any* way *categorize* blood-sucking, mating, and avoiding being bitten in two, as ways for things to go better in its life. It has no notion of the good, and little time for contemplation of the good's several manifestations. But no such thing is needed to make sense of the tick's actions as serving a better-making purpose in its life.

To sum up, I have characterized agency as follows.

1. Agency is a capacity for one distinctive type of activity among others, possessed by some living things among others.
2. Agency is a *mindful* or *aware* capacity to *guide* the trajectory of one's own life and the events in one's environment …
3. … where the 'guidance' is *shaped* by the agent's discriminations between different features of the world …
4. … and where the 'guidance' is thus neither *invariant* nor *random*, but is rather geared towards bringing it about that things go better, precisely by dint of the impact that agency has on the situation.

III.

At the start of this paper, I said that avoiding anthropocentrism in action theory gives us new insight into the *norms* of action. Let me now explain that claim. First, the foregoing characterization of agency allows us to evaluate particular exercises of agency (and particular agents) in more than one dimension. To illustrate the rich range of evaluations permitted by this conception of agency, let us consider the case of Oedipus.

Oedipus, King of Thebes, killed his father, married his mother, violated the precepts and values most dear to him, and despaired. Let's imagine that Oedipus was simply as good as it gets at *human* modes of agency. Let's stipulate that he did not betray a "tragic flaw" when he killed Laius on the road. Let's pretend that the ancient Greek moral code of honor, courage, and nobility was entirely correct, that Oedipus adhered to it absolutely, and that modern values of non-violence, tolerance, and diplomacy are mistaken. If so, then in his life, Oedipus acted well. Indeed, he exhibited a *remarkable* degree of perfection of human agency. In short, Oedipus was wise. Neither when he married his

mother, nor when he loved his children, nor when he killed his father on the road, nor when he tirelessly sought the true murderer of Laius, nor when he refrained from suicide in his despair, did he make the kind of error where he failed to use the tools of action available to human beings as well as he might have.

And yet Oedipus failed spectacularly to *interact* well with certain highly salient features of the way his world 'really is'. That is, through blameless, fated ignorance, he failed to take account of some of the facts. This failing was a practical failing, clearly: the interactions that he initiated with the world did not guide the trajectory of events in a way that was warranted by the situation. At the same time, his missteps, which ruined his life, were in themselves (by hypothesis) due to *no imperfections in his exercise of human agency*.

Equipped with the characterization of agency we've been discussing, we can explain this tragic bind. The failure of Oedipus to achieve human success and well-being shows that human practical perfection is not equivalent to, and does not guarantee, perfectly measuring up to the norms of agency. Exercising agency (acting) is one kind of activity, with certain success standards. Humans have our own ways of engaging in this activity: we have language, memory, opposable thumbs, etc.. But just like all the other kinds of agents in the world, human beings as a species may be equipped with imperfectly sophisticated and imperfectly effective *forms* of agency. We can ask whether the way *humans* characteristically do 'being agents' is a good/the best way engage in aware, organism-initiated interactions with the world. We can ask, for instance, whether *rationality* is well-suited to realize the purpose or role it has developed to serve in human life. Are rational agents effective, good, accurate inter-actors with their environment? In which respects are they better and worse? When a person leads a happy human life, largely constituted by wise actions and fortunate experiences, are they leading a better life than a cockroach or a super-human being could live? Could one sometimes be *wiser* by being *less rational*? And so on. We can ask these sorts of questions with perfect intelligibility, once we have a sense of the characteristics of agency that is independent of how they appear in the human case.

Once we conceive of agency in more general terms, we may also be led to acknowledge certain respects in which other kinds of agents have an edge on us, for instance with respect to the accuracy, fruitfulness, sensitivity, efficiency, or sustainability of their mode of interaction with our shared environment. For instance, *migration* is of great interest to human beings, because migratory birds and butterflies navigate the globe with great precision, again without tools (and even apparently without memory, since they do it, in the case of Monarch butterflies, over more than one generation). It's a live question whether their way(s) or our way of getting from point A to point B is more sophisticated, more efficient, more accurate, more reliable, etc., as compared to our technologically assisted alternatives.

In closing – the case of Oedipus underlines the 'world-answerable' character of action, according to the characterization of agency that I've been sketching. Oedipus based his actions and decisions on his best assessment of how the world really is, and of what in the world is worth choosing and going for, and why. In just the same way, our evaluations of Oedipus' actions are based on *our* assessments of how his (fictional) world really is, including how things 'really' were with him, as well as what in the world is worth choosing and going for, and why. These evaluations and assessments and decisions and choices are all experience-based, and they can all be wrong in part or in full. They can all be revised in light of further experience. We can – indeed in the case of Oedipus, people probably will– spend the rest of time arguing about whether a given exercise of agency – a given action – was better or worse all things considered, or in some particular point of detail. In some very difficult cases, we might never know for sure. Nonetheless, recognizing action as an interaction between world and agent makes actions answerable to the world. It thus promises to properly situate not only human agency, but also the norms and standards that pertain to human action, in the broader context of animal agency and animal success and failure, where it belongs.

References

Anscombe, G.E.M. *Intention*. Oxford: Basil Blackwell, 1957.

———. "On Brute Facts." *Analysis* 18 (1958): 69-72.

Aristotle, *Nicomachean Ethics.* Trans. Christopher Rowe. Oxford: Oxford University Press, 2002.

Burge, Tyler. *Origins of Objectivity.* Oxford: Oxford University Press, 2010..

Frankfurt, Harry. "The Problem of Action". Reprinted in *The Importance of What We Care About*. Cambridge: Cambridge University Press, 1988.

Geach, Peter. "Good and Evil." *Analysis* 17 (1956): 33-42.

Lawrence, Gavin. "The Function of the Function Argument." *Ancient Philosophy* 21 (2001): 445-75.

Mackie, J.L. *Ethics: Inventing Right and Wrong*. New York: Penguin, 1977.

McDowell, John. "Might There Be External Reasons?" In *World, Minds, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*. Edited by J.E.J. Altham and Ross Harrison. Cambridge: Cambridge University Press, 1995: 387-98.

———. *Mind and World*. Cambridge, Massachusetts: Harvard University Press, 1996.

Murdoch, Iris. "The Idea of Perfection." In *The Sovereignty of Good*. London: Routledge, 2010.

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

Yujia Song
University of North Carolina at Chapel Hill

Yujia Song is a graduate student at the Philosophy Department of UNC - Chapel Hill, primarily interested in normative ethics.

"Empathy, Proper empathy, and Understanding"

Abstract*: The aim of this paper is to explore the proper place of empathy in morality, and in the course of doing that, bring out the relationship between empathy and understanding. I will argue that empathy plays two roles in morality: first, empathy itself can constitute the appropriate moral response in a situation; and second, quite independently from its first role, empathy can contribute to our forming an appropriate moral response by supplying us with information that we can use to understand the situation and the people involved in it. Those interested in empathy tend not to recognize that both of these two roles are limited. In particular, with regards to the second role, they tend to mistake what is in fact a failure in understanding for a failure to exercise empathy properly.*

1. Introduction

The aim of this paper is to find a proper place of empathy in morality, and in the course of doing that, bring out the relationship between empathy and understanding. I will argue that there are two major roles empathy plays in morality:[91] first, empathy itself can constitute the appropriate moral response in a situation; and second, quite independently from its first role, empathy is an important means for understanding the people involved in a situation, thus indirectly enabling us to form an appropriate moral response.[92] Those interested in empathy tend not to recognize that both of these two roles are limited. In particular, with regards to the second role, they tend to mistake what is in fact a failure in *understanding* for a failure to exercise empathy properly.

I will begin by introducing two promising attempts at finding a proper place for empathy in morality. Both aim to carve out a notion of "proper empathy," for empathy, if unconstrained, can go wrong in different ways. I will then give a diagnosis of these accounts and explain why they are unsatisfactory. Following that, I will draw two lessons from these failed attempts. The first lesson is that part of what we want "proper empathy" to do can indeed be done by empathy, and that is its first role I have mentioned above. However, the other part of what we want "proper empathy" to do – and this is the second lesson -- cannot, and should not, be the job of empathy, but of understanding.

2. Why empathy unconstrained is not good enough

Before I begin, just a few words on what I mean by "empathy." Although usually taken to be a *feeling* congruent to another's feeling or situation, there is large consensus that empathy, as a way to vicariously experience another's inner states, involves both an affective and a cognitive aspect. More than feeling how the other person is feeling, empathy is often a matter of "putting oneself into another's shoes," imagining not just her feelings, but her beliefs, desires, perceptions, and so on. There is much debate over how we go about taking up another's perspective, but suffice to say, insofar as empathy involves perspective-taking, it has both affective and cognitive components.

---

[91] I do not think that these two roles exhaust all the work empathy does in morality. One might plausibly think, for example, that a capacity for empathy is necessary for one's capacity for being moral at all.

[92] Following Laurence Blum, I take it that giving a moral response to a situation is not only a matter of determining what the correct action is (and carrying it out), but it also involves first "apprehending" the situation. That is to say, even though the situation is given, *what it is a situation of* is still up to the agent. How the agent responds to the situation first of all depends on how she apprehends the situation.

The way philosophers have attempted to locate empathy in morality has been much influenced by the close connection between empathy and altruistic behavior. C.D. Batson's studies that purport to prove the "empathy-altruism hypothesis" is often cited by philosophers.[93] According to the hypothesis, the feeling of empathy gives rise to the motivation to promote another's well-being purely for that person's sake. Nancy Sherman suggests that we need to cultivate empathy in order to cultivate altruistic virtues like benevolence. Alisa Carse takes it one step further, arguing that properly cultivated empathy itself is a moral virtue. Michael Slote goes even further. Assuming that empathy is essential to caring, Slote revamps the ethics of care with a focus on empathy. On his view, empathy properly exercised not only can serve as the standard by which an act is judged right or wrong, but grounds the other key notions in morality such as justice and rights.

But wherever they want to put empathy in morality, philosophers are all alarmed by some pretty serious problems posed by empathy if it is not properly done. I will sketch two main problems here.

First, as Hume observes, and as empirical studies have confirmed, we tend to empathize more with people we are more similar to, or closer to in terms of physical distance or personal relations. But if empathy is to ground our altruistic response, or more generally, any kind of moral response, we must not follow our natural tendencies that bias our empathic engagement, for proper moral response calls for at least a reasonable degree of impartiality.

Another problem has to do with maintaining one's sense of the boundaries between the self and the other. The two extremes of empathic involvement – too much focus on the self, or on the other -- are each termed "incuriosity" and "self-effacement," in Carse's terminology.[94] Carse distinguishes between two types of "incuriosity." The first results from *too little* identification with the other, due to indifferent or negative attitudes towards that person for various reasons. Alternatively, Carse suggests, one may project her own experience onto the other in an attempt to empathize, if there is too much identification such that one fails to recognize the difference between her own experience and that of the other person. The other extreme, "self-effacement" or "self-denial," occurs when one lets the other's feelings, thoughts, desires, and so on overtake one's own, thus compromising one's integrity.

3. Introducing proper empathy

Since empathy done "improperly" cannot deliver the promises that philosophers hope it will, there is a strong temptation to "proper-ize" empathy, to put it under certain constraints. I want to focus on two such proposals, one by Carse and the other by Slote. Carse wants to put empathy under *moral* constraints, whereas Slote seems to appeal to *psychological* standards for his conception of proper empathy.

Carse constructs what she calls "morally contoured empathy – empathy properly felt and expressed (p.171)." The idea is that moral principles should be incorporated into a conception of "proper empathy" to correct for problems like incuriosity and self-effacement. They can be impartial principles and considerations, which urge the agent to step back and observe the situation from a detached, objective point of view. More importantly, Carse thinks, empathy should be guided by "normatively substantive conceptions of our roles and relationships and their defining moral stakes (p.171)." These relationship-based principles allow the agent to remain in a stance of attachment, and yet empathize in accordance with the expectations, responsibilities and obligations that arise from her role or her relationship with the other person.

---

[93] Although as Stephen Darwall points out, Batson's notion of "empathy" is closer to what we would mean by "sympathy," i.e. concern for another as we apprehend her situation from our own point of view. There has been much confusion, since Hume's discussion of the subject, over the meaning of two terms, as well as the connection between empathy and sympathy.

[94] Carse's characterization of the two poles of "improper empathy" is in turn adapted from the two poles of "imaginative involvement" identified by Piper.

Slote comes to a conception of proper empathy in a different way. He claims that "[t]he way to correct morally misguided or inadequate empathy is not, I believe, with new and different mechanisms or procedure, but with more or more thoroughgoing empathy (2010, p.52)." If one is fully developed in terms of feeling empathic concern for others, then her empathic engagement would be *just right*, morally speaking. It's not clear what Slote means by "fully developed empathic concern." All he says is that it is "the kind of empathy that would exist in human circumstances favorable to the overall development of empathy (2007, p.30)." The emphasis on development and his particular focus on Martin Hoffman's studies in the development of empathy suggest that he may be talking about an ideal in some *psychological* sense: we can extrapolate from the progress that children make as they get better at empathizing to get some idea of what a person *fully* developed in this capacity would be like.[95]

Slote does not elaborate on what this ideal kind of empathic concern amounts to. I gather from his book and his responses to the critics that someone who exhibits fully developed empathic concern exhibits the following features: (a) shows concern for *every* person who is involved in the situation or will be affected by the agent's action; (b) is able to correct our natural tendency for similarity or familiarity bias; (c) is able to exercise partiality towards those she is closer to where appropriate;[96] (d) maintains a proper balance between being empathically engaged with others and retaining her own capacities, interests, thoughts, and feelings.

4. Diagnosis of accounts of proper empathy

The problem with these attempts at proper-izing empathy is that they confound different normative standards at work. To correct for all the ways in which empathy could go wrong, we need at least these three sets of standards[97]:

(1) Constitutive norms: norms internal to empathy, that dictate how successful one is at empathizing with another. The closer one is at observing these norms, the more successful one is in her attempt to empathize. And when one completely violates the norms, she can be said to be not empathizing at all.

Thus, it is a norm internal to empathy that one should have a fairly accurate idea of the other's inner states (given objective limitations). If one takes her own response to the other's situation to be what the other is experiencing, instead of imagining how the other would respond given that person's particular background, personal traits, etc., then one is barely following this norm. She therefore fares poorly at empathizing.[98] And in cases where one's idea of the other's experience is utterly inaccurate, it seems that we can no longer say she is empathizing with the other.

(2) Psychological norms: norms that determine whether one is psychologically healthy and fitting for her developmental stage. I do not have a full account of what such norms encompass, but I will mention a few that are relevant to empathy. One norm has to do with self-other differentiation. As Martin Hoffman notes in his description of the development of empathic distress, children develop from having no clear sense of the self-other distinction, to recognizing the separateness of the self and others, and to realizing that each person has her own inner states. Thus, a mature, psychologically

[95] Yet at the same time Slote might also be thinking of a moral ideal, for the cultivation of empathy may well incorporate the guidance of moral principles. But given his overall project, this cannot be the case. The reason is quite clearly that if he is to ground other moral notions in the notion of "fully developed empathic concern," then he cannot invoke the former in his analysis of the latter without going in a circle. Therefore, it must be empathy fully developed in some *non-moral* sense.

[96] While Slote wants to be able to account for obligations to people beyond our own group, he also wants to account for a kind of partialism he endorses.

[97] I do not claim that these three are the *only* norms at work, but I think they are the most relevant to the present discussion.

[98] However, projecting our own thoughts and feelings onto the other person does often work well even though it is not strictly speaking the same as taking up the other person's point of view, since we do have much in common.

healthy person is able to perceive herself as distinct from other people, with a distinct set of thoughts, feelings, desires, goals and values. And she recognizes the same goes for other individuals. The upshot of this point is that if one is confused about which inner states belong to herself or to others, or even does not realize that she has a separate set of inner states, she is not psychologically healthy.

Related to the self-other differentiation, and perhaps even overlapping somewhat, is what Adrian Piper terms "unity and rational integrity of the self" (734). This involves preserving one's desires, thoughts, values, actions, and so on, against interference from the outside, and maintaining some level of coherence in them. One way the integrity of the self can be compromised is through a complete identification with someone else, where the other person's inner states or experience are taken to be one's own and subsequently drown out the set of thoughts, feelings, etc. that makes up the self. This is also a breakdown of the self-other distinction.

Additionally, a psychologically healthy person has the capacity for empathy (in both the affective and cognitive respects), that is, the capacity to go beyond one's own thoughts and feelings, and to experience what it is like to be the other person in that person's situation. This capacity need not be exercised all the time (well, maybe it *should* not be). There are times when we think that one "ought to" exercise this capacity -- such as when one's friend is grieving for the death of a family member -- but it would not be a *psychological* failure if the person does not in fact empathize (although if someone never shows empathy even at occasions that typically trigger empathic response, we have reason to suspect that the person is incapable of empathy).

(3) Moral norms: without committing ourselves to any particular normative ethical theory, we can still agree on some moral norms, such as we should help the needy, or that parents should take care of their children. We may also agree on certain moral principles. Impartiality, for instance, requires that the agent consider the interests of all parties involved in a situation rather than pick and choose as she likes.

I want to make two points about the distinction between different norms. First, the norms "internal" to empathy are not moral norms. We're concerned here not with whether one should empathize, or how much one should empathize, but with how good she is at carrying out the task of empathy. Someone who empathizes well, meaning who is highly accurate in her awareness of what the other is feeling and thinking, may or may not meet relevant moral standards. A father who does an impeccable job empathizing with his son's craving for junk food may nevertheless come to endorse it and neglect his responsibility as a father to help the child establish a healthy diet (and develop self-control).

Secondly, although violations of the "internal" norms or psychological norms need not always also be violations of moral norms, often times they are. And when they are, the moral failure may be a *consequence* of a failure in empathy or a failure to maintain one's psychological wellbeing. One may fail to respond to another's plight because, having failed to empathize accurately with the other, she forms a wrong idea of what the other is going through. This is an example of a violation of moral norms due to a violation of "internal" norms of empathy. We can find an example of moral failure due to a violation of psychological failure in the wife who submits completely to her husband's wishes and values at the expense of giving up her own. Here there is no violation of the "internal" norms of empathy, for she can see perfectly, albeit too perfectly, from her husband's perspective, and understands perfectly well his inner states. But her empathy with her husband is inappropriate at the psychological level, for she loses herself in the total identification with her husband. And from a moral

point of view, this undermining of her integrity constitutes a failure to respect herself, to see herself as being equal to her husband, and her own thoughts, feelings, and needs no less valuable than his.[99]

The point is that empathy can be done "improperly" in different senses. Sometimes one may be empathizing accurately, but still considered to be "improper" because her response to the situation or to other people is *morally* inappropriate. Sometimes one neglects to empathize with those that one has a moral obligation to. We call this too an instance of improper empathy, but there is in fact in this case a *lack* of empathy.

One problem with Carse's and Slote's accounts of proper empathy is now clear. Both are oblivious to the different kinds of norms at work. In the case of Carse, the two poles of empathy – "incuriosity" and "self-effacement" – are judged improper under different sets of norms. Self-effacement is considered improper not because empathy itself is inaccurate, but because it threatens one's psychological well being, or compromises one's capacity as an autonomous moral agent. As for incuriosity, the two ways by which it can manifest also fall under different kinds of norms. The kind of incuriosity due to indifference or prejudice is considered improper because one fails to empathize; the failure is a moral failure. The other kind of incuriosity, however, does not denote a lack of empathy, but empathy exercised inaccurately as one projects her own thoughts and feelings on to the person she intends to empathize with. Here, the internal norms of empathy are violated. Whether or not any moral norms are violated as well is quite a contingent matter.

Compared to Carse, Slote makes a more serious mistake. Not only does he also confuse these different kinds of norms, but he is wrong to think that his conception of proper empathy relies solely on non-moral (perhaps psychological) norms. Towards the end of section 3 I sketched out a rough picture of what proper empathy amounts to for Slote. The basic idea is that one must have empathy with the right people and to the right degree – not too much or too little relative to everyone else in the situation and to oneself. But moral norms have clearly been smuggled into this picture, for the internal norms of empathy or psychological norms cannot get us as far as being able to determine the "right people" or the "right degree." What is "right" is relative to what is appropriate moral response to a situation, and that is determined by moral norms.

Why is it important to point out this confusion over norms? So that we can be aware of two distinct concerns philosophers have. On the one hand, they want to claim that in certain situations, it is (morally) proper for the agent to *have empathy*. On the other hand, they also want to claim that if an agent does empathize, it is (morally) proper for her to do it *in the right way*. I will argue in section 5 that the proper place for empathy in morality has to do with the value of empathy itself, independently from whether or not one empathizes well (i.e. accurately). But the second concern, as I will argue in section 6, is misguided. What philosophers are really interested in, when they try to direct empathy "in the right way," is sufficient *understanding* of the situation and the people involved in it. But if this is their real goal, empathy alone will not get them there, no matter how well it is exercised.

5. Restoring empathy's place in morality

I think empathy is morally significant by virtue of the fact that it allows us to step outside our own point of view, directing our attention from the self to the other.[100] The significance of the non-

[99] What about a wife who wholeheartedly agrees with her husband's goals and values such that she adopts them as her own and actively takes upon herself to promote them? I'm hesitant to say that in this case, she has "lost" herself in merging with her husband. What's the difference between this woman and one who *does* exemplify a case of "self-effacement"?

[100] This is not to say that in empathizing with someone, we already feel *for* her or are moved to promote her good *for* her sake. Stephen Darwall makes it clear that at the stage of empathy, we are taking the other's perspective, putting ourselves *in* her position, and sharing the experience *of* her, but we are not concerned *for* her yet, as we would be at the stage of sympathy (p. 263-364). It is therefore a little misleading to say that we direct our attention from the self to the other when we empathize. The point is really that as we try to empathize, we direct our attention away from what concerns ourselves to what concerns the other person *as she would attend to it*. It is also not to say that empathy is the only means by which we

egoistic feature of empathy lies partly in its contribution to altruistic behavior. The claim is not a very strong one. It is neither that empathy necessarily leads to altruism, nor that altruistic behavior requires empathy[101]. Stephen Darwall distinguishes empathy from sympathy by suggesting that there is not yet a (feeling, action, etc.) "*for* the other" in empathy although one does engage with the other's inner states. But precisely because empathy directs our attention away from the self, it puts us in a position where we can readily respond to the other, in ways that are congruent with her needs and interests, for her sake.[102]

A second point I want to make about the non-egoistic feature of empathy builds upon Carse's idea that empathy is constitutive of caring relationships. Carse suggests that empathy is valuable (and indeed, indispensible) in such relationships "not only as a crucial epistemic aid, but also *intrinsically* (188)…" I take her to mean that empathy is valuable regardless of how well one empathizes with the other, or how well one adheres to relevant moral principles in empathizing. There's something valuable about empathizing itself, and I think it has to do again with the non-egoistic stance one adopts in empathy. It enables one to reach out to the other and establish a sense of connection.[103]

6. Empathy and understanding

"Proper" empathy is also important for its contribution to the agent's understanding of the situation by giving her insight into the experience of those involved in it. Independently of whether empathy is part of the proper moral response in a situation, it can still supply the agent with information to shape her response. The constraints that Carse, Slote and others put on empathy are to ensure that empathy plays both roles properly; first, that it is exercised, and second, that it is exercised such that the agent reaches a pretty good understanding of the people involved. But while we are

---

direct our attention away from the self to the other. Again, the distinction between empathy and sympathy is useful here. When one feels sympathy for another person, she is focused on that person's well-being, not her own. She need not take up the other's point of view, i.e. empathize with the other. But like in empathy, she steps outside of her personal concerns.

[101] It is not required in the sense that one can act for another's good in the absence of empathy. But as Sherman points out, it is indeed required in the sense that one needs the knowledge about the other person obtained through empathy so as to succeed in one's attempt at doing good for the other's sake.

[102] Why, then, think that caring *requires* empathy? I think empathy supplies two important elements. First, unlike some other altruistic behavior, caring requires that the one-caring establishes and expresses a sense of connection (or attachment?) with the one cared-for. This is similar to the point I'm about to make regarding the role of empathy in close personal relationships. Second, like other altruistic behavior, caring requires that the one-caring have sufficient knowledge about the one cared-for and her situation to provide the kind of care that is needed.

[103] Immediately following the claim I just quoted, Carse goes on to illustrate her point with an example, one in which a mother is oblivious to her partner's abuse of her daughter. She writes:

> "As the mother, it would, to be sure, be crucial that I 'wake up' and see what is going on in my home. Empathic imagination might, as I have suggested, be essential to achieving this urgent epistemic demand. But simple awareness is not enough. What is also vitally needed is felt comprehension of my daughter's suffering, an emotional resonance *that conveys to her* that I grasp, or am attempting to grasp, the enormity of the psychological injury she has endured through the abuse *in its meaning for her* – in this case, perhaps, her sense of isolation, fury, despair, or revulsion. In the absence of empathy of this kind, it is unlikely that a connection with her can be restored or that trust can be repaired. A failure of empathic engagement would, in this case, be tantamount to egregious abandonment. It would not suffice as proper maternal care."

Although Carse clearly stresses the importance of sufficient *understanding* (what is epistemically demanded of the mother) of what the daughter is going through, her main concern here is that the mother's act of empathy itself, or attempt at it, is crucial for the relationship and for her role as a mother. Without engaging with her daughter empathically, Carse says, the mother would be unable to restore a *connection* with her daughter, and she would be guilty of "egregious abandonment" of her daughter. The emphasis falls on the sense of connection or attachment that is created and maintained when the mother is empathically engaged with the daughter. The mother effectively conveys to the daughter that she feels with her, stands alongside her. It is not the same as feeling bad for her (in sympathy), or uncritically endorsing how she sees everything from her point of view.

concerned with empathy *itself* when we consider its first role, what we are actually interested in regarding its second role is not empathy itself, but understanding.

Why do I say that our interest in proper empathy is an interest in understanding in disguise, and hence better served by directly addressing issues in understanding? Let us look again at what the agent must do to count as achieving proper empathy:

First, the agent must reach a reasonably high level of accuracy in grasping the inner states of the other. Improper empathy occurs when one misunderstands what the other thinks or feels, thereby violating the norms *internal* to empathy. And since a major source of empathic inaccuracy lies in too much attention to the self, we are to guard against our tendency to impose our own experiences on others. "Self-absorption," as Carse calls this phenomenon, is seen as a huge threat to having proper empathy. To overcome it is to check one's prejudices or assumptions about the other, to refrain from projecting one's own values or expectations on to the other, and essentially, to get at certain *facts* about the inner states of the person one is trying to empathize with. That is to say, proper empathy aims at *understanding* of the experience of the other.

Moreover, to properly empathize in any situation, according to both Carse and Slote, the agent cannot arbitrarily pick whom to empathize with. Rather, she should empathize with *all* who are involved in the situation. In one example that Carse discusses, the mother who is too engrossed with her partner fails to notice he is abusing her daughter. On Carse's view, the mother fails to properly empathize with her daughter as she is oblivious to the latter's struggles and pain, even though she may be empathizing with her partner well enough (perhaps even "too much"). This failure is twofold. As Carse puts it, the mother fails to (i) "resonate in feeling or imagination with [her daughter]," and (ii) "be curious about the emotions she is expressing in her conduct (p.178). I think these two parts of the verdict map nicely onto the two roles of empathy outlined earlier in the section. The lack of empathy (with her daughter) on the mother's part is at once a failure of appropriate moral response to the daughter, and, at least as Carse sees it, an important factor in the mother's distorted view of what is going on in her family and her subsequent failure to take action against her partner. One may object that the mother does not need to *empathize* with the daughter to find out about the abuse. She can be more *observant* of the behavior of her partner and of her daughter. She can also talk to her daughter about how things are going for her. The goal is to understand what is happening to the daughter so as to have a full, accurate view of the situation the family is in, although doing these may lead the mother to empathize with her daughter (which in turn enhances understanding). Slote, on the other hand, has in mind a different set of examples, but what I have to say is quite similar. He suggests that one who empathizes properly is able to extend empathy beyond those we are close to or familiar with. Such an agent is better able to respond to a situation where different parties – some closer to her than others – have a stake in it but with conflicting interests. But here again, the problem is not that we don't empathize enough with certain people, but rather our understanding of the situation is flawed.

While "too little" empathy is no good, "too much" of it can also be a problem. Self-effacement can be harmful to one's integrity, and in other cases, undermine one's ability to fulfill the responsibilities she has towards the person she is empathizing with. Carse seems to claim that in such cases, the correct response would require the agent to limit the extent of her empathic engagement with others and maintain a kind of distance and independence from the objects of her empathy. However, I think the problems associated with self-effacement have not to do with empathy itself, but with the way the agent uses the knowledge she has gained through empathy. Where the agent is deemed to have "too much" empathy, what she is doing is in fact relying too heavily (even exclusively) on empathy to get at an understanding of the situation. While empathy does supply useful information about the other,

the heavy emphasis on empathy can also preclude the agent from coming to a fuller understanding of the other (by way of adopting alternative perspectives, for example).[104]

So far, I have argued that our interest in proper empathy derives from an interest in understanding. But this would not be a problem for those advocating a notion of proper empathy if empathy is *the* way to understand others. Yet, I think empathy is neither sufficient nor most reliable for understanding. As we have seen in the above discussion of self-effacement, one's grasp of the other's experience through empathy gives one an incomplete and one-sided view of the other. And since one takes the point of view of the other while empathizing, one may be blind to the actual condition of that person if the latter happens to have a rather distorted picture of herself without realizing it.[105] Empathy tells us what it is like to think or feel like the other person, but it alone does not also inform us of *what it is* for one to think or feel that way. It is only when one steps back from empathic engagement and reflects on the vicarious experience that she can come to understand the person more accurately.

Bibliography

Baier, Annette C. (2010). "Is Empathy All We Need?" *Abstracta*, Special Issue V: 28-41.

Blum, Lawrence A. (1980). *Friendship, Altruism and Morality*. London ; Boston: Routledge & Kegan Paul.

Carse, Alisa L. (2005). "The Moral Contours of Empathy." *Ethical Theory and Moral Practice: An International Forum*, 8(1-2), 169-195.

Cottingham, John. (2010). "Empathy and Ethics." *Abstracta*, Special Issue V: 13-19.

Darwall, Stephen. (1998). "Empathy, Sympathy, Care." *Philosophical Studies* 89 (2-3): 261–282.

Driver, Julia. (2010). "Caring and Empathy: On Michael Slote's Sentimentalist Ethics." *Abstracta*, Special Issue V: 20-27.

Hoffman, Martin L. (2000). *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge, U.K.; New York: Cambridge University Press.

Noddings, Nel (2010). "Complexity in Caring and Empathy." *Abstracta*, Special Issue V: 6-12.

Piper, Adrian M. S. (1991). "Impartiality, Compassion, and Modal Imagination." *Ethics*, Vol. 101, No. 4. pp. 726-757.

Sherman, Nancy. (1998). "Empathy and Imagination." *Midwest Studies in Philosophy*, 22:1. 82 – 119.

Slote, Michael (2007). *The Ethics of Car and Empathy*. London and New York: Routledge.

--. (2010). "Reply to Noddings, Cottingham, Driver, and Baier." *Abstracta*, Special Issue V: 42-61.

[104] I want to note that one who suffers from self-effacement also has deficient understanding of *herself*, which in turn adds to a faulty judgment about how one is to respond to the other.

[105] The image of a person with mental illness comes to mind when we think of someone with a "distorted" picture of herself. I do not wish to restrict the class of people to those afflicted with mental illness. One may be unaware of certain facts about herself, or self-deceived about them, without being mentally ill. Limited or partially false self-understanding is common, or even an inescapable human condition.

Keynote Address:

Harry G. Frankfurt

Princeton University

Harry G. Frankfurt is professor emeritus of philosophy at Princeton University. He was a member of the Department of Philosophy from 1990-2002. His books include *Demons, dreamers, and madmen*; *The defense of reason in Descartes's Meditations*; *The Importance of What We Care About*; *Necessity, Volition, and Love*; *The Reasons of Love*; *On Bullshit* ; and *On Truth*.

"Volitional Rationality and the Necessities of Love"

1. According to some philosophers, moral judgments are susceptible to decisive rational justification, or to decisive rational rejection, entirely on the basis of unimpeachably objective considerations. They insist that moral judgments must not be understood as depending essentially, for warranted affirmation or denial, on any considerations that are ultimately subjective – as, for instance, facts concerning a person's attitudes or desires, or facts concerning how a person feels. They maintain, in other words, that it is possible to establish the acceptability or the unacceptability of a judgment of practical reason without appealing at all to the occurrence or to the absence of any subjective state of mind.

Among the philosophers who have recently defended this anti-subjectivist position are Philippa Foot and Michael Smith. They recommend their own objectivist alternatives to moral subjectivism, in part, by issuing a warning: if the claims of moral objectivism are not adopted, and we are required to accept a subjectivist account of morality, the position in which we will then find ourselves is manifestly intolerable.

2. Foot maintains bluntly that the subjectivist account cannot possibly be correct: "it just can't be . . . ," she declares, "that morality in the end is just the expression of an attitude. . . . [If that were what it is, then] whatever reasons might be given for a moral judgment, people might without error refuse to assent to it, not finding the pertinent feelings or attitudes in themselves. . . . [T]here is no way, if one takes this [subjectivist] line, that one could imagine oneself saying to a Nazi, 'but we are right, and you are wrong' with there being any substance to the statement. Faced with the Nazis, who felt that they had been justified in doing what they did, there could simply be a stand-off. And I thought: 'Morality just cannot be subjective in the way that different attitudes, like some aesthetic ones, or likes and dislikes, are subjective.'"[106]

As for Smith, the threat of moral subjectivism appears to arouse in him an even more disturbing anxiety. He warns of "the panic that we rightly feel when we reflect upon the possibility that we can give no privileged rational defense of moral concern . . . ."[107] To protect ourselves from experiencing this panic, he believes, we must maintain the conviction that human reason enjoys a legitimate and effective authority to evaluate moral judgments objectively.

3. Foot and Smith are evidently driven to their views by the same concern. They worry that if a subjectivist account of moral judgments were correct, it would be necessary to concede that there is

---

[106] *Harvard Review of Philosophy*, vol. xi, Spring 2003, p. 34, emphasis added.

[107] "Dispositional Theories of Value," *Mind*, --------, p. 103, emphasis added.

no such thing as a genuine moral reality, independent of the subjective vagaries of the human mind, against which the truth or falsity of a moral judgment might objectively be measured. In that case, moral disputes could not satisfactorily be resolved on the impersonal grounds that reason provides. The acceptability or the unacceptability of moral judgments could only be determined by considering the feelings and attitudes by which various people happen disparately to be moved.

People are notoriously heterogeneous, of course, regarding what they like or dislike, what they are inclined to allow or to prohibit, what appeals to them or what they find appalling. The dictates of reason, on the other hand, are the same for everyone: whether a certain argument is valid or invalid, whether a certain proposition is or is not self-contradictory, whether some empirical judgment conforms or fails to conform to the relevant facts – these matters are independent of what anyone (or, indeed, of what everyone) may think or feel about them.

Foot and Smith believe that if morality lacks an objective foundation, there can be no way to provide universally authoritative justifications of moral judgments. By the same token, there can be no decisive way to justify condemning opponents whose moral judgments are based on opposed attitudes or feelings. Both Foot and Smith are troubled fundamentally by the thought that if moral subjectivism were correct, it would be rationally unjustified for us either wholeheartedly to endorse a moral claim or to regard the denial of that claim as being unequivocally a mistake.

Thus, we could not reasonably maintain that the moral views of even our most viciously inhumane enemies are erroneous. Additionally, we might well be unable to avoid panic when we realized that we cannot provide even our own moral convictions with satisfyingly objective rational support. The whole enterprise of moral inquiry and advocacy would collapse, then, into an inchoate morass of idiosyncratic sentiment and propensity.

4. Foot calls attention especially to the alleged importance of our having been warranted to contend that the moral views of the Nazis were not right but wrong. Now, it goes without saying that Nazi morality was horrifyingly depraved, and that we were certainly justified in undertaking to expunge it from our civilization. But was it really essential, as Foot evidently believes it was, for us to have considered the moral beliefs of the Nazis to be erroneous, or wrong?

Surely, what was most important was not to condemn or to refute the moral beliefs of the Nazis. It was to fight the Nazis, and to defeat them. It is not clear, indeed, that it was important at all for us to believe that the moral views of the Nazis were mistaken. For in fact, reasonable and adequate grounds of quite another sort were available for regarding those views as hateful. That is, we may be objectively justified in being vigorously opposed to certain people without supposing that any of their beliefs is a mistake.

If someone were to attack one of my beloved children, there would be general agreement that I would be justified in seeking to defend my child against the attack. This general agreement that such a response on my part would be justified does not rest upon any presumption that the attacker would have made a mistake. To begin with, the fact that I love my child entails quite strictly that, when it seems to me that the child is in danger, I will be moved powerfully to protect it. Next, consider the fact that lovers identify with the interests of their beloveds, regarding those interests as their own. Thus, offering proportionate resistance to an attack on the well-being of what one loves is justifiable – other things being equal -- as being, in effect, an act of self-defense.[108]

The trouble with the Nazis was not that, with regard to issues of morality, they had made a mistake. Nor was it necessary for us to have believed that they were in error in order for us reasonably to have considered ourselves justified in aiming to destroy them. The most conspicuously alarming

[108] Of course, this reason for defending my child may be overridden by other considerations. While it justifies my *being moved* to defend the well-being of the child, it leaves open the possibility that there may be superior justification for me or for others to prefer an alternative course of action.

trouble with the Nazis was that they threatened irreparable injury to something -- our culture and our ideals -- which we love.

5. Foot is skeptical that a subjectivist morality can account for such transparently reasonable and compelling moral intuitions as that it was justified for us to stand up to the Nazis. I believe that her doubts are due to an erroneous presumption concerning the kinds of subjective states upon which subjectivist theorists necessarily consider moral judgments to rest. Foot presumes, mistakenly, that subjectivists invariably consider moral judgments to rest upon such states as those to which she refers as "likes and dislikes" – in other words, upon attitudes and feelings that may be quite shallow and transient, and that may merely happen casually to pass through our minds.

As Foot correctly insists, it would be preposterous to suppose that morality rests upon nothing more solid, or more steadily authoritative, than such ephemera. In my own subjectivist view, however, it is not at all on that sort of state that morality is grounded. What I maintain is that the subjective state on which moral judgments are appropriately grounded is, rather, the state of love.

So: what is love? The conception of love I propose does not aim at encompassing every feature of the hopelessly disordered set of conditions that people commonly think of as instances of love. Most especially, the phenomenon I have in mind is not to be confused with infatuation, dependency, obsession, lust, or similar varieties of psychic turbulence. As I construe it, love is a particular mode of caring. It is a non-voluntary, non-utilitarian, rigidly focused, and – as is any mode of caring – a self-affirming concern for the flourishing of what is loved.

The object of love can be almost anything – a kind of experience, a kind of activity, a person, a group, a moral ideal, a non-moral ideal, a tradition, whatever. The lover's concern is rigidly focused in that there can be no equivalent substitute for the loved object, which the lover loves in its sheer particularity and not as an exemplar of some general type. The lover's concern is non-utilitarian in that the lover cares about the beloved for its own sake, rather than only as a means to something else.

It is an essential feature of love that the lover identifies with the beloved, thus taking the interests of the beloved as his or her own. This identification is non-voluntary, in that it is not under the immediate control of the will. A person cannot love – or stop loving – merely by deciding to do so, or by judging that it would be desirable to do so. Parmenides said that love is "the first-born offspring of necessity."[109] We come to love because we cannot help loving. Love is a non-rational condition. It requires no reasons, and it can have anything as its cause

Genuine love is certainly not a mere casual impulse, nor a momentary flicker of sentiment or inclination. To be sure, it may not be unshakably permanent. On the other hand, people possess no capacity to initiate it or to dispel it merely by a simple exercise of will. Love is neither a product of voluntary choice nor is it nothing more than just a bit of whimsy. It is deeply grounded in a person's non-voluntary predilections – i.e., in what the person is determined, by his own nature, to care about in a certain way; and it possesses a native stability and endurance.

Moreover, by virtue of the fact that love entails commanding requirements and constraints on the lover's behavior, it brings its own necessities and its own authority. Since it defines the limits of the lover's will, it determines thereby the shape of his or her volitional identity. Thus love, at least as I understand it, is very far from being either transient or shallow.

6. Against the view I have enunciated -- that morality is grounded in love -- Smith argues that what we love is too contingent to support the impersonal commitment required by morality. He speaks of the citizens of Leningrad who, during the Second World War, accepted horrendous sacrifices in defending their city against the Nazis. He suggests that, although it might indeed have been because they loved their city that they accepted those sacrifices, love of their city would not have sufficed to justify what they did. The reason it would not have sufficed, Smith argues, is that the people of

[109] J. Burnet, *Early Greek Philosophy*, (London, 1948), p.177, fragment 13.

Leningrad could not have provided any satisfactory rational justification for loving Leningrad. Of course, they could have accounted for their love of Leningrad by referring to the fact that they had grown up there. But they would have had to acknowledge that they would probably have loved Berlin, and not Leningrad, if instead they had grown up in Berlin. In Smith's view, the contingency of what we happen to love makes it impossible to regard love as providing sufficient authority to ground a binding commitment.

We must surely presume that morality is impersonal, and that its scope is universal. It seems reasonable to suppose, then, that we cannot consider the authority of morality over an individual to depend upon, or to be derivative from, particular contingent features of that person's circumstances. Whatever an individual's contingent personal circumstances may happen to be, the person is necessarily bound to comply with what morality universally requires. So, just as it is clear that a person cannot elude those requirements in virtue of any contingencies, it equally cannot be supposed that contingencies of any kind are their source or their ground.

However, the insistent authority of love's requirements is not truly undermined by the fact that what a person loves is a contingent matter. The people of Leningrad might not have loved that city, of course, if they had grown up elsewhere. But that does not show that their love of Leningrad was inadequately authoritative.

After all, if the circumstances of my life had been different, I might not have come to love the woman and the children to whom – by virtue of my love for them -- I am now actually bound. Acknowledging this possibility surely does not mean that I must regard my present commitments to my wife and my children as less binding on me than they would have been if no other commitments had been possible.

My devotion to my wife and to my children, which is entailed by my love for them, is immune to the fact that I might not have had that wife or those children. A commitment is no less compelling because it is merely a contingent fact that it has been incurred. Thus, similarly, it is obvious that a person's obligation to fulfill a promise is not undermined by the fact that the person might not have made that promise.

The necessities entailed by a particular love are indeed only contingent necessities, since it is only a contingent fact that an individual happens to be bound by that specific love. But they are necessities nonetheless, from which the individual cannot escape merely at will. They bind sufficiently, then, to satisfy the condition that the commands and constraints of morality must be rigorously unavoidable.

7. But it is not enough to point out that contingent love imposes necessities that, like the necessities imposed by morality, cannot voluntarily be eluded. There remain the glaring facts that people do not all love the same things, and that some loves strike us as less rational than others. These facts may appear to imply that, if morality were to be based on love, the authority of the moral law could not be truly universal or rational.

For it may seem that what people love is so various and so indiscriminate that this cannot help but infect -- with a severely unacceptable relativism -- any morality that proposes to derive its essential support from love. Given the variety of what people love, how can we suppose that what we happen to love provides any "satisfactory rational justification for loving" it?

I am inclined to think, however, that there are in fact certain things that all people love, and that they cannot help loving. In my view, moreover, it is precisely our love of those things that provides the source and the ground of our moral understandings. It has become customary to be excessively attentive, it seems to me, to the variegated genetic and environmental determinants that account for what often appears to be a radical distinctiveness of individual lives. Those determinants are widely presumed to ensure that there is no such thing as a common human nature, and that each of us is unique.

In certain respects, perhaps, we are indeed all different. It may well be that, insofar as we are genetically and environmentally molded humans, no one of us is exactly the same as any other. But while the variety of our social and cultural environments may ensure that there is no common human nature, which each human being invariably shares, we do all share a common nature as human animals.

There are, in fact, a variety of final goals – ends that move us in themselves -- to which all of us are, by virtue of our nature as animals and regardless of the particular circumstances of our lives, innately committed. We are not interested in these things merely as means, which we think will be helpful in enabling us to attain other things. They are final ends, which we desire for their own sakes. Moreover, we cannot help doing so; our desires for them are non-voluntary and stable elements of our nature. They may appropriately be regarded, then, as being objects of our love. We do not love them because we have reasons for loving them. We love them just because that is how we are made.[110]

8. Here are some of the ends-in-themselves which, by our very nature as self-conscious animals, we all seek and love: each of us naturally seeks and loves to avoid bodily harm, physical and psychological agony, failure in what we undertake to accomplish, deprivation of what we need, and death; nor can we tolerate extended personal isolation, constriction of our mobility, or prolonged empty boredom. Furthermore, we have a natural empathy for others, that leads us -- in one degree or another -- to care that these evils be avoided in their lives as well as in our own.

Avoiding each of these universally unwelcome conditions is something to which we are by nature devoted. It is something that we naturally love, and that we cannot help loving. Each is a goal whose pursuit, at least in part, defines us and delineates our true interest.

It goes without saying that circumstances may arise in which we would be willing, or even eager, to accept one or another of these naturally hated evils. Under certain conditions, a person may agree to serious deprivation, or may accept death, or may voluntarily endure some other of the normally unwelcome conditions to which I have referred. In such cases, however, what is desired or accepted is not loved. It continues to be counted as undesirable; but in the circumstances, it counts as less undesirable than some other condition, for the avoidance of which it is regarded as an indispensable means. It is not considered to be an end-in-itself, which is valued entirely for its own sake. It is not desired, then, as an object of love.

What we regard as the moral law essentially consists of an elaborated and refined codification and prioritization of rules designed to facilitate the realization of these goals.[111] This means that it is, after all, possible for a moral principle to be objectively either correct or mistaken: for, as a matter of objective fact, a principle or rule either actually does tend to maximize the realization of these final ends, or it does not.

This is far from implying, however, that morality is thoroughly objective. For the goals themselves, from which the constitution of morality essentially derives and on which it is grounded, are ends-in-themselves just because we love them. That is, they are our final goals simply by virtue of certain subjective states – i.e., by virtue of certain inclinations, feelings, and attitudes.

9. There remains, then, the critical question of whether morality can be justified, finally and rationally, by providing an acceptable basis for considering our love of these ultimate goals to be itself rational. For it is one thing to maintain that moral judgments can be evaluated objectively by measuring them against goals that are loved by everyone; but it is clearly another thing to maintain that

---

[110] Of course, there may be important genetically or environmentally conditioned differences in the *intensity* with which they are loved and in the *order of priority or preference* each of them is assigned.

[111] The elucidation and justification of this claim I leave as an exercise for the reader.

it is rationally justifiable for us to love those goals in preference to others. In other words, we must consider whether it may be that morality -- despite, or even on account of, being grounded fundamentally on what we all naturally love – is basically not rational.

A productive approach to dealing with this issue may be found, I believe, in considering Hume's contention that the dictates of morality are decidedly not rational. Hume makes his point by way of a famous example, which illustrates his view that we have no rational basis for our ultimate final ends. Even the most grotesque preferences, he insists in his example, are not irrational. More specifically, he says that "'tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger."[112]

It is true that a preference for destruction of the world involves no purely logical mistake. So far as logic alone is concerned, it is unobjectionable: someone who chooses to protect his finger from a trivial injury at the cost of unlimited destruction elsewhere is not thereby guilty of a contradiction, or of a faulty inference. In this purely formal sense of rationality, his choice is not at all irrational.

We would nonetheless, of course, condemn that choice. But what would we actually say of someone who embraced it? Surely we would not merely complain against that person, as Foot would have wished us to complain against the Nazis, that we are right and that he (or she) is wrong. What we would surely say is, rather, that the person must be crazy. In other words, despite the unassailability of his (or her) preference on strictly formal grounds, we would consider both it and him (or her) to be wildly irrational. Caring more about a scratched finger than about "destruction of the whole world" is not just an error, or a blunder, or an unappealing quirk. It is lunatic. Anybody who has that preference is a madman.

10. When we characterize the person in Hume's example as "crazy," or as a "lunatic," or as "mad," these epithets do not function merely as vituperative rhetoric. They are to be understood literally, as denials that the person is a fully rational creature. There is a rather familiar mode of rationality, then, which is not essentially defined by a priori, formal necessities. Hume's lunatic may be fully competent in constructing valid chains of inference, and in distinguishing between what is and what is not logically possible. Indeed, the irrationality in question is not fundamentally a cognitive deficiency at all. The person is volitionally irrational. He (or she) has a defect of the will, which bears on what are his (or her) final ends and thus on how he (or she) is disposed to choose and to act. It concerns what the person loves.

In leading us to consider the person to be volitionally irrational, the critical point has to do with possibilities: the lunatic is prepared to implement voluntarily a choice that we ourselves could not bring ourselves to make. There are notable structural analogies or parallels between the contingent necessities of volitional rationality and the strictly formal a priori requirements of pure reason. Both modes of rationality limit what is possible; and each imposes, accordingly, a corresponding necessity. The boundaries of formal rationality are defined by the necessities of logic, to which no exceptions are conceivable. The boundaries of volitional rationality, on the other hand, are defined by the necessities of love.

The latter do not effectively constrain conceptual functions or capacities. Rather, they constrain the will. They limit what it is in fact possible for us to care about or to take as ends, what we can accept as reasons for action, and what we can actually bring ourselves to do. Circumstances that violate the volitional necessities of love, unlike circumstances that violate the limits of logical possibility, are not inconceivable. What stands in the way of violations of love is, instead, that they are unthinkable.

---

[112] David Hume, *A Treatise of Human Nature*, edited by L.A. Selby-Bigge, Oxford University Press, 1888, Book II, Part III, section III, p. 416 (emphasis added).

11. It should be noted that being volitionally rational is not a matter just of the choices a person actually makes or is inclined to make. More fundamentally, it involves an inability to make certain choices. If a person undertakes to reach a cool and balanced judgment concerning whether it would be a good idea to destroy the entire world in order to avoid being scratched on a finger, that does not manifest a sturdy rationality. Even if the person were finally to conclude that destroying the world to protect the finger is not a good idea, the fact that it was necessary to deliberate about this makes it clear that there is a serious deficiency in what the person loves.

Rationality does not permit us to be open-minded, or judiciously deliberative, about everything. It requires that certain choices be utterly out of the question. Just as a person transgresses the boundaries of formal reason by supposing that some self-contradictory state of affairs might really be possible, so a person transgresses the boundaries of volitional rationality by regarding certain logically possible alternatives as genuine options.

People who are volitionally rational cannot bring themselves to choose or to do various things which, so far as power and skill alone are concerned, they are entirely capable of choosing or doing. They are subject to a volitional necessity, by which their wills are bound within certain limits; thus, there are certain courses of action which they are constrained from deciding effectively to pursue. They may think that a certain course of action would be appropriate, or even that it is mandated; but, when the chips are down, they cannot go through with it. They cannot mobilize the will to implement the judgment. Destroying the world to avoid a scratch on a finger is something volitionally rational people cannot bring themselves to do. In virtue of the volitional necessities of love, by which their wills are bound, making that choice – or pursuing that alternative -- is not genuinely among their options. It is simply unthinkable.

What makes it unthinkable? Why are we unable to bring ourselves to make certain decisions or to do certain things? What accounts for our inability, or for our inflexible refusal, to include among our live options various alternatives which we are otherwise quite capable of pursuing? What is the ground of the constraints on our will that volitional rationality entails?

One view is that these volitional necessities are responses to an independent normative reality. On this account, which is endorsed by such philosophers as Foot and Smith, certain things are inherently important. By virtue of their inherent importance, they provide reasons for acting in certain ways. This importance, and the fact that they provide reasons, is not a function of our attitudes or beliefs or desires, or of subjective factors of any kind. It does not depend on what we are actually inclined to accept as reasons for acting, or in any way on the character or condition of our will. Rather, it is alleged that these reasons possess an inescapable normative authority. It is the natural authority of the real, to which all rational thought and conduct must seek to conform.

Advocates of objectivist moral theory ordinarily leave it very unclear by just what observation or procedure the independent reality of these reasons is supposed to become apparent to us. They generally simply presume that we do somehow recognize, with vivid clarity, that various things are inherently important. Once we have done that, we cannot help accepting the authority of the reasons these things provide. It is impossible for us to hold back from acknowledging the importance that is – so to speak – right before our eyes. After all, seeing is believing. Thus, what imposes constraints on our will is just the incontrovertibly forceful immediacy of reality itself.

This is the doctrine of "normative realism". It holds that there are objective reasons for us to act in various ways, whether we know them, or care about them, or not. If we fail to appreciate those reasons, we are making a mistake. The widespread prevalence of the belief that we need to avoid error in our normative judgments and attitudes – that we need to get them right – is often cited as providing in itself weighty support for the view that the importance of reasons is inherent in them and that practical reason is therefore securely grounded on the independent reality of its governing norms.

12. My own view is different. I do not believe that anything is inherently important. In my judgment, normativity is not a feature of reality that is altogether independent of us. The standards of volitional rationality, and of practical or moral reason, derive -- so far as I can see – only from ourselves. They are not grounded, of course, on what we merely happen to like or to desire. Rather, they are grounded on what we cannot help caring about and thus on what we cannot help considering to be important.

This necessity bears pertinently, I believe, upon the question of whether we can find a "satisfactory rational justification for loving" what we love. The fact that we cannot help loving what we love means that we have no choice but to love it; and if we have no choice, then there is not much point in asking whether our love is justifiable. We will go on with it, after all, regardless of whether or not it can be justified.

To be sure, there is nevertheless still a sense in which we may reasonably wish to consider whether our love is rational. For we may wonder whether, despite its inescapability, our love may be irrational in the sense of being self-defeating or destructive. What would make it irrational in this way would be the incompatibility of its necessities – the reasons it engenders -- with the reasons engendered by some other of our loves.

Thus, there are two elements in a satisfactory rational justification for loving what we love. If in fact we cannot help loving it, then whether the love is rationally justifiable does not genuinely arise as a practical question. And if loving it is also coherent with the necessities of our other loves, then there can be no basis for considering it to be irrational.

13. Our judgments concerning normative requirements can certainly get things wrong. There is indeed an objective moral reality, which is not up to us and to which we are bound to conform. However, this reality is not objective in the sense of being entirely outside our minds and decisively independent of us. Its objectivity, and its independence of us, consist just in the fact that it is outside the scope of our voluntary control.

Normative truths require strictly that we submit to them. What makes them inescapable, however, is not that they are grounded on an external reality. They are inescapable because they are determined by volitional necessities that we cannot alter or elude. In matters concerning morality, and concerning practical norms more generally, the objective reality that requires us to keep an eye out for the correctness of our views is a reality that is within ourselves.

Edward S. Hinchman
University of Wisconsin Milwaukee

Edward Hinchman teaches at the University of Wisconsin-Milwaukee and works on issues in epistemology and moral psychology. His recent publications in moral psychology include "Receptivity and the Will" (*Noûs*, 2009); "Conspiracy, Commitment, and the Self" (*Ethics*, 2010); and "Narrative and the Stability of Intention" (*European Journal of Philosophy*, forthcoming).

"Rational Requirements and 'Rational' Akrasia"

Abstract: Can akrasia be rational? Can it be rational to resist the motivational force of your own practical judgment? While I don't believe that akrasia can be rational, I think there's something revealingly right in the idea. The fundamental issue lies in the relationship between two conceptions of rationality. Previous treatments of 'rational' akrasia – for example, Harry Frankfurt's – have regarded rationality as a responsiveness to reasons. Previous treatments of rational requirements have regarded rationality as an attitudinal coherence. I'll argue that rational requirements codify an agential coherence that you negotiate through a dynamic of self-trust and self-mistrust. It is not reasoning to abandon your judgment through forgetfulness, confusion or perverse self-rebellion. But, contra some assumptions informing T. M. Scanlon's influential account of practical judgment, it can be reasoning to abandon your judgment through reasonable self-mistrust. The difference lies in how self-mistrust can manifest a sensitivity to the norm of coherence that gives force to rational requirements.

On one conception of practical rationality, being rational is most fundamentally a matter of avoiding incoherent combinations of attitudes. This conception construes the norms of rationality as codified by rational requirements, and one plausible rational requirement is that you not be akratic: that you not judge, all things considered, that you ought to φ while failing to choose or intend to φ. On another conception of practical rationality, being rational is most fundamentally a matter of thinking or acting in a way that's informed by your practical reasons. This second conception construes the norms of rationality in terms that appear to allow the possibility of rational akrasia, since your capacity to act on your reasons can function at a level that need not involve deliberative judgment.[113] Though their treatments of akrasia make them seem incompatible, I'll argue that the two conceptions of rationality are not incompatible. It is possible to accommodate the core insight motivating defenses of 'rational' akrasia within the conception of rationality as codified by requirements of rational coherence.

On the conception of rationality as codified by requirements of rational coherence, talk of 'rational akrasia' is self-contradictory. But I'll argue that defenses of 'rational' akrasia simply misformulate their core insight. What those arguments aim to vindicate is better understood as a thesis about how you can reason your way out of akratic irrationality. The obvious way to do it is by bringing your inclinations to choose or intend into line with your practical judgment. But the insight that talk of 'rational' akrasia misformulates is that you can also reason your way out of akratic irrationality by bringing your practical judgment into line with your inclinations to choose or intend.

[113] As Harry Frankfurt and others have emphasized, it is plausible to regard your deliberative judgment as merely one medium for registering your practical reasons, with emotions and undeliberated habits serving as other media.

My principal aim is to explain how an agent can do that compatibly with the conception of rationality as codified by requirements of rational coherence. As we'll see, abandoning your practical judgment in response to a disinclination to commit yourself to it can be a way of *reasoning* your way from akratic irrationality into rational coherence. Rational requirements codify not logical or formal coherence, I'll argue, but an agential coherence that you negotiate through a dynamic of self-trust and self-mistrust. It is not reasoning to abandon your judgment through forgetfulness, confusion or perverse self-rebellion. But it can be reasoning to abandon your judgment through reasonable self-mistrust. The difference lies in how self-mistrust can manifest a sensitivity to the norm of rational coherence that gives normative force to rational requirements. The core insight of those who defend the possibility of 'rational' akrasia lies in their emphasis on the rational force of self-mistrust.

Here's an overview of how I'll argue for this rapprochement between the two conceptions of rationality. I'll set terms for the discussion in section I, briefly sketching Harry Frankfurt's defense of the possibility of 'rational' akrasia, then rehearsing a more recent debate between John Broome and Niko Kolodny about the nature of the rational requirement that akrasia violates. I'll accept the debate's guiding conception of rational requirements as specifying relations of rational coherence among an agent's judgments and attitudinal commitments. In the example we'll work from, the coherence is between a practical judgment and the 'downstream' attitude, choice or intention, whereby one commits oneself to executing the judgment in action. The question is how this requirement of rationality governs the 'direction' of reasoning. In section II, I'll make a proposal that Broome and Kolodny do not properly consider: that akratic incoherence – a self-conscious failure to commit to your own judgment – could provide a basis for reasoning in an 'upstream' direction to the abandonment of that judgment.[114] I'll argue that the question of rationality in cases of 'rational' akrasia addresses not the akratic state itself but the agent's commitment to resolve it toward rational coherence. If we can understand how one might make this resolution in an 'upstream' direction – abandoning the judgment *because* one akratically fails to commit to it – we'll have preserved a core strand in the idea that akrasia could be rationally praiseworthy. I'll reply to an objection in section III by diagnosing why one is bound to misconstrue the idea of upstream reasoning if one accepts, as both Broome and Kolodny do, T. M. Scanlon's account of practical judgment, an account that my argument gives grounds for rejecting. Once we clarify the nature of rational coherence, we'll see how the ancient virtue of rational self-control – *enkrateia* – can counter akrasia in either direction. What previous discussions treat as 'rational' akrasia is simply your predicament when enkrateia requires you to abandon your judgment rather than follow through on it.

I.

Why might one regard akrasia as rational? Though other authors have discussed the phenomenon more recently and in greater detail,[115] we can work from Harry Frankfurt's seminal treatment in "Rationality and the Unthinkable."[116] Frankfurt presents a case from Trollope in which Lord Fawn can't bring himself to implement his own best judgment that he ought to consult an eyewitness about

[114] I take the stream metaphor from Kolodny ("Why Be Rational?," *Mind* 114 (July 2005), 509-563; and "State or Process Requirements?," *Mind* 116 (April 2007), 371-385), though like any metaphor it can mislead. The stream flows from judgment to choice or intention. But in another sense the only 'flow' here is the process of your reasoning. Of course, when you reason 'upstream' you are not at all swimming against *that*! Talk of 'reasoning' can also mislead: we are not assuming that all reasoning involves the deliberative weighing of reasons. A non-metaphorical way of putting the thesis will emerge in sections III through V.

[115] In the full version of this paper, I discuss work by Nomy Arpaly and Karen Jones at length.

[116] In his *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988).

the questionable conduct of his fiancée. As Frankfurt describes the akrasia, "[i]t is Fawn's judgment that the best thing for him to do would be to speak with Gowran about Lizzie, and he tries to carry out that course of action. At bottom, however, he is unwilling for his will to be shaped in that way."[117] Frankfurt argues that Fawn's akrasia does not necessarily render him irrational, since whether akrasia is irrational depends on whether the feelings that counter the agent's judgment track his reasons. "A person's judgment may itself be radically contrary to reason," Frankfurt argues: "it may well be that a failure of his will to accord with his judgment is precisely what saves him from irrationality."[118]

Though there is much that is plausible in Frankfurt's argument, we may nonetheless wonder about the idea that akrasia can be rational. His akrasia puts Fawn in violation of what appears to be a core requirement of rationality: that you not judge, all things considered, that you ought to φ without also committing to that judgment by choosing or intending to φ. Isn't it flatly irrational to persist, even for a moment, in judging it best to φ while resisting that judgment – your own judgment – by failing to choose or intend to φ? If that state is irrational – as I'll concede it is – how could moving 'upstream' from the failure to choose or intend to φ the abandonment of the judgment that you ought to φ, as Frankfurt describes Fawn as doing, be itself anything but irrational? If such an upstream maneuver is irrational, what becomes of the compelling intuitions that drive Frankfurt's description of the akratic transition as engaging the very essence of rationality, insofar as it enjoins us not merely to be narrowly coherent but to act from reasons?

I'll approach these questions via a debate between John Broome and Niko Kolodny about the normative structure of rational requirements. The debate turns on two questions: first, whether rational requirements have wide scope or narrow; second, whether they are state requirements or process requirements. We can illustrate with the requirement that Lord Fawn violates (which is the requirement most discussed by both authors): that you intend to do what you judge you ought to do, all things considered. In narrow-scope formulation, that would read:

Narrow Necessarily, if you judge that you ought to φ, then rationality requires that you intend to φ.

And in wide-scope formulation, it would read:

Wide Necessarily, rationality requires that (if you judge that you ought to φ, then you intend to φ).

On **Wide** your judging that you ought to φ would not generate a requirement that you intend to φ, but on **Narrow** it would. On **Wide**, rationality requires only that you either form the intention or cease to make that judgment. Perhaps you ought to do the latter. Broome argues that to avoid illicit bootstrapping we must regard rational requirements as having wide scope.[119]

---

[117] "Rationality and the Unthinkable," 183.

[118] "Rationality and the Unthinkable," 189.

[119] See Broome "Normative Requirements," *Ratio* 12 (December 1999), 398-419; "Normative Practical Reasoning," Proceedings of the Aristotelian Society, Supplementary Volume 75 (2001), 175-193, at 181-2; "Are Intentions Reasons? And How Should We Cope with Incommensurable Values?" in C. W. Morris and A. Ripstein (eds), *Practical Rationality and Preference: Essays for David Gauthier* (Cambridge: Cambridge University Press, 2001), 98-120; "Practical Reasoning," in J. Bermùdez and A. Millar (eds), *Reason and Nature: Essays in the Theory of Rationality* (Oxford: Oxford University Press, 2002), 85–111, at 92-7; and "Reasons," in R. J. Wallace, P. Pettit, S. Scheffler and M. Smith (eds), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (Oxford: Oxford University Press, 2004), 28-55, at 29. Broome also gives other arguments for the thesis in the earlier of these papers. But when he defends it in more recent papers – e.g. in "Does rationality consist in responding correctly to reasons?," *Journal of Moral Philosophy*, 4 (November

Broome's defense of the wide-scope reading dovetails with his position on the second question: rational requirements require merely that you not be in the state of having undertaken the commitment while failing to follow through on it.[120] If rational requirements are state requirements, thus defined, then you could satisfy a rational requirement – that is, do what rationality requires of you – simply by abandoning the commitment. But the consequent of that conditional is the claim that Kolodny rejects. The only way to satisfy a rational requirement is to be guided by it, in Kolodny's metaphor, 'going forward' through a process of 'downstream' reasoning. In this respect, he concludes, rational requirements are process requirements.

In his main argument for this conclusion,[121] Kolodny observes that the scope debate would not seem pressing if the requirements at issue were state requirements. Here are the requirements formulated as narrow- and wide-scope state requirements:

Narrow State	Necessarily, if you judge at t that you ought to φ, then rationality requires that you intend at t to φ.
Wide State	Necessarily, rationality requires that you not be in the following state: you judge at t that you ought to φ, but you do not intend at t to φ.

When we apply these principles we confront three possibilities: you make the judgment and form the intention, you make the judgment but do not form the intention, or you do not make the judgment. Kolodny observes that **Narrow State** and **Wide State** yield different results only in the third case: you satisfy **Wide State** but merely fail to violate **Narrow State**. Therefore, "the choice between wide and narrow scope matters only in so far as the difference between satisfying a requirement and not violating it matters."[122] But that is not, Kolodny notes, an important difference in the context of his debate with Broome.

We can reframe the question of 'rational' akrasia if we follow Kolodny on this point. Since we do find the scope debate pressing, let's treat it as a debate over process requirements. Here is a narrow-scope process requirement:

2007), 349–74, at 354 – he uses only the argument from illicit bootstrapping. In "Normative Requirements," Broome calls this a 'non-detaching' normative relation (as opposed to narrow-scope 'detaching' relations). The terminology of 'narrow scope' and 'wide scope' appears for the first time in Broome's 2004 paper "Reasons." As Broome acknowledges, the worry about bootstrapping generalizes an argument made earlier by Michael Bratman (*Intention, Plans, and Practical Reason* (Cambridge: Harvard University Press, 1987), 24-7).

[120] The distinction between 'state' and process' requirements was coined by Kolodny ("Why Be Rational?," 517) in the course of arguing against Broome. Broome subsequently ("Wide or Narrow Scope?," *Mind* 116 (April 2007), 359-370, at 366) embraced the thesis that rational requirements are state requirements.

[121] This 'main' argument is in fact Kolodny's third argument against Broome on this point. Kolodny first observes that "our ordinary attributions of irrationality are at least sometimes about what people do, or refuse to do, over time" ("State or Process Requirements?," 371), which suggests that at least some of the requirements at issue are process requirements. He next observes that at least some of these requirements "can function as advice or guide your deliberation" (ibid., 371-2), which appears to conflict with Broome's claim that they are state requirements, since a state requirement tells you merely to avoid a certain state without telling which of the two ways of doing so to pursue. These arguments fail to do much damage, however, since Broome can reply that avoiding or exiting from an irrational state is, first, something that people do or fail to do over time and, second, something that you can advise someone to do, where the content of the advice is disjunctive. There is nothing in general problematic about disjunctive advice, or about the idea that one is guided by disjunctive advice. True, you cannot determine what to do if that is all you go by. But there is no reason to think that the rational requirement is all the subject has to go by.

[122] "State or Process Requirements?," 374.

Narrow Process Necessarily, if you judge at t that you ought to φ, but you do not intend at t to φ, then rationality requires you to form going forward from t, on the basis of the content of your judgment, the intention to φ.

And here is Kolodny's formulation of the cognate wide-scope process requirement:

Wide Process$_K$ Necessarily, if you judge at t that you ought to φ, but you do not intend at t to φ, then rationality requires you (EITHER to form going forward from t, on the basis of the content [of] your judgment, the intention to φ, OR to revise going forward from t, on the basis of the content of your lack of an intention to φ, your judgment that you ought to φ).

Kolodny now argues against **Wide Process$_K$** by claiming merely that its second disjunct "makes no sense," since "[y]our lack of an intention to [φ] has no content."[123]

II

One problem with Kolodny's argument against the wide-scope interpretation of process requirements is that his formulation, **Wide Process$_K$**, manifests an assumption that he does not defend: that the process governed by a rational requirement must be a process of reasoning from propositional contents. As I'll now argue, we can plausibly reject that assumption.

To see the problem with Kolodny's assumption, let's formulate **Wide Process** without it:

Wide Process Necessarily, if you judge at t that you ought to φ, but you do not intend at t to φ, then rationality requires you (EITHER to form going forward from t, *on the basis of your judgment*, the intention to φ, OR to revise going forward from t, *on the basis of your failure to intend* to φ, your judgment that you ought to φ).

The revised requirement has the virtue not only of making actual sense but of capturing the real basis of your reasoning in both disjuncts. In the first disjunct, you don't form the intention to φ on the basis of *the content* of your judgment that you ought to φ, as **Wide Process$_K$** puts it, but on the basis of *your judging* (or of your *having judged*) that you ought to φ. The content of your judgment is merely a proposition. What leads you to form the intention is not that proposition but your judgment with that proposition as its content. This would be a mere quibble were it not for the fact that Kolodny's argument depends on this aspect of his formulation. His formulation obscures how you can – 'going forward' in your reasoning – come to abandon your judgment that you ought to φ on the basis of your

[123] "State or Process Requirements?," 379. (In both **Wide Process$_K$** and in this quoted passage, Kolodny has "*X*" instead of "φ.") Kolodny goes on (ibid., 379-81) to consider the idea that you might reason upstream from *other believed contents* in the neighborhood of your failure to intend to φ. But, as he rightly notes, if that is how we construe the second disjunct of our **Wide Process**, and if you have no such other beliefs, then all we have left is the first disjunct, and our **Wide Process** collapses into **Narrow Process**.

failure to form an intention to φ.[124]

You can do so because the failure is not a mere lack. You do not reason from your mere lack of an intention to φ, or from (as Kolodny unintelligibly puts it) the 'content' of this lack. You reason from an aspect of your *failure* to form the intention. Calling it a 'failure' acknowledges the bearing of a norm on your agency at this juncture. And the norm is indeed just the rational requirement under discussion, about which we are wondering whether it is best captured by **Narrow Process** or by **Wide Process**. So your failure, insofar as you fail to intend to φ while judging that you ought to φ, is the failure to satisfy this rational requirement. It is a breakdown in rationality, and the question for you is how to respond. Here is one way you might respond: unform your judgment that you ought to φ.

Let me quickly forestall some misunderstandings. Most crucially, my proposal is not that you reason either from the proposition *that* you judge that you ought to φ or from the proposition *that* you fail to intend to φ. I propose that you reason either from the judgment itself (with its propositional content) or from your specific failure – retaining that judgment – to form the cognate intention (with its content). The propositional contents of the judgment and of the intention are obviously a big part of what drives the reasoning, but another part of what drives the reasoning is the judgment itself or the failure to form the intention itself. We might say that the propositional element drives the reasoning via its participation in the attitudinal element. Just as the attitudes are propositional, so the propositions are attitudinal: they figure as the contents of attitudes the forming and failing to form which can amount to transitions that constitute stages in a process of reasoning. Reasoning proceeds *through* such transitions in attitude; it is not as such *about* transitions or attitudes.[125]

How then do the attitudes themselves (with their contents) drive the process of reasoning? My proposal is straightforward: through the relational attitudes of self-trust and self-mistrust.[126] I'll

---

[124] I pursue this point toward a challenge to Kolodny's error-theoretic treatment of the normativity of rationality in "Reasons and Rational Coherence," in preparation. But that challenge need play no role in the argument we're developing here.

[125] Of course, one can also reason about one's attitudes. But then the transitions that constitute the process of reasoning have second-order contents, which is not the case we're considering. (This point will become important to my argument in section VI.)

[126] The restriction to self-mistrust explains what's wrong with Kolodny's treatment of the possibility that one might reason in an 'upstream' direction. Consider this passage:

> One can reason from the content of one's belief that one lacks sufficient reason to *X* to dropping one's intention to *X*. One can rationally resolve the conflict in this way. But one cannot reason from the content of one's intention to *X* to revising one's belief that one lacks sufficient reason to *X*. It is not reason to cling to what one judges to be an unfounded intention and to support it by revising one's belief about one's reasons. It is a kind of wishful thinking or self-deception. Consider, to a first approximation, how one would express this transition to oneself: 'The facts of my situation do not give me sufficient reason to *X*. I hereby commit to doing *X*. Thus, all along, the facts of my situation gave me sufficient reason to *X*.' I say 'to a first approximation,' because the 'thus' seems out of place. ("Why Be Rational?," 528-9)

There are several problems here. First, this is the core of Kolodny's argument against 'upstream' reasoning, and he discusses only the possibility of intending to φ while judging that you lack sufficient reason to φ. The cases that interest us, by contrast, are ones in which you judge that you have sufficient reason to φ – that is, that you ought to φ, all things considered – without intending to φ. I do not claim that it is possible to intend to φ while judging that you lack sufficient reason to φ. In the cases of 'rational' akrasia on which we'll continue to focus, there's a failure to intend as you judge best. If there's also intending as you fail to judge best, that's a further case, or a further aspect of the case, and not one that I plan to discuss. Second, in claiming that such an 'upstream' transition must manifest wishful thinking or self-deception, Kolodny overlooks the possibility that I'm emphasizing: that the transition is grounded in self-mistrust. Third, if we stipulate that the transition is grounded in self-mistrust and that it reflects merely a failure to intend, we can explain the 'thus' in your reasoning by noting that you could make your reasoning explicit like this: 'I judge that the facts of my situation give me sufficient reason to φ. But I can't bring myself to trust that judgment by forming a practical commitment to φ. Thus, [here you abandon the judgment] it is not the case that the facts of my situation give me sufficient reason to φ.' The reasoning proceeds through a tension in your thinking that is indeed rationally incoherent, since the 'thus' expresses

explain the proposal at length in the rest of the paper, but let me emphasize two key aspects of it immediately. First, self-mistrust is a rich intrapersonal attitude and relation. When I say that you fail to intend as you judge through self-mistrust, I don't mean merely that you fail to form the intention. You might fail to form the intention in any number of ways: for example, through distraction, forgetfulness, confusion, perverse self-rebellion, or an interruption of your consciousness caused by bumping your head. In none of these cases would it make sense to say that you fail to intend as you judge because you mistrust this deliberative judgment. The failure to intend as you judge that I have in mind is specifically a failure that we can explain (or at least begin to explain) in terms of self-mistrust.[127] Second, intrapersonal attitudes of self-trust and self-mistrust need not be, and typically are not, mediated by judgments that you are trustworthy or untrustworthy. You don't typically form an intention to φ because you judge that you are trustworthy in judging that you ought to φ, and mistrusting your judgment that you ought to φ need not involve the judgment that you are untrustworthy in so judging.[128]

Such cases are not uncommon. You're single and deliberating whether to invite your new romantic interest away for the weekend. Or you're wondering whether to invest in a certain stock. Or whether to make a sudden career move. You reach the conclusion that you should go for it, that indeed you should make the decisive call right now. But as you reach for your phone you pull back. Does this irrational failure to commit to your judgment manifest nothing more than fear? If so, you'd be a coward to let it stand in the way. But it may mark a different attitude. It may manifest a sense that you're simply not a good judge on this matter – at least, here and now. For a few moments your self-doubt makes you akratic: you judge that you ought to make the call but also resist that judgment. As you doubt yourself further, however, you find yourself wondering whether making the call is really such a good idea. That is, you find yourself redeliberating the matter. Your self-doubt has thus undone your judgment, but not by pointing to any reasons that you'd earlier overlooked. The self-doubt comes first and *causes* you to take another look at your reasons. Again, the transition could manifest second-order judgments about your own trustworthiness. But that it now how it proceeds in a typical case, and (as we'll see from a different angle in section III) there are philosophical confusions in the idea that such a transition must involve second-order thinking.

Let me emphasize once again that while gripped by self-mistrust of the sort that I'm characterizing – that is, while mistrusting a judgment that you've made and retain – you are in an irrational state. The question is whether a rational *process* can take you through such a state. That the state is irrational shows nothing about the rationality of the process. Once we grant that judgment-to-commitment akrasia is possible, we concede that any time you form a judgment that you've not yet committed to you are in the same irrational state. It doesn't follow that you cannot reason self-trustingly to an undertaking of the commitment. If we think you can, then we can ask whether that also goes for the other direction: can you rationally resolve your irrational state by letting self-mistrust unform your judgment? We might regard the two processes as, rationally speaking, on a par: in some cases, the rational process is to form the commitment self-trustingly; in other cases, the rational process is to unform the judgment self-mistrustingly. In which direction should you now reason?

---

your mistrustful rejection of the authority of your own act or attitude of deliberative judgment. But the 'thus' also resolves the tension: since you can't bring yourself to form a commitment to φ, you abandon your judgment that you ought to φ. This is the proposal that I'll defend – and of course clarify further – in what follows.

[127] By 'self-mistrust' I don't mean mistrusting yourself in general but merely in the specific instance at hand. A more general attitude of self-mistrust may emerge, but it need not. And I'm not talking about mistrusting your judgment *sans phrase*, but about mistrusting a specific judgment you've made on a specific matter at hand. By 'trust your judgment,' I'll always mean *trust a specific judgment*, never *trust your faculty of judgment*.

[128] I'll devote the entirety of section III to diagnosing confusions in the literature on this and related points. For now, let me simply stipulate that the relations of self-trust and self-mistrust that I'll be discussing need not be mediated by judgments of trustworthiness or untrustworthiness.

Well, that depends on whether your judgment is from your own (possibly non-deliberative) perspective worthy of your trust.[129]

III

One might nonetheless worry that self-mistrust can figure only in downstream reasoning. Cannot an akratic recalcitrance to your own judgment count as evidence that you should not trust that judgment? Say you regard yourself as having concluded deliberation with the judgment, all things considered, that you ought to φ. But now, thinking further, you feel the force of self-mistrust. Why not regard that further thinking as an extension of your deliberation, wherein your mistrust reveals that 'all things' were not considered, even by your own lights, when you regarded yourself as drawing that deliberative conclusion? To deal with this worry we must consider more fully the model of practical judgment presupposed by Broome and Kolodny. We can thereby diagnose why they failed to anticipate my defense of **Wide Process**. And we can dispel the worry by seeing why we should reject that model of practical judgment.

Broome and Kolodny share a conception of practical judgment as importing a species of commitment about which one could reason only downstream. Underlying the conception are two more specific assumptions, both of which derive from an influential account of judgment pioneered by T. M. Scanlon.[130] First, this conception assimilates practical commitment to a species of doxastic commitment, since it equates your all-things-considered practical judgment that you ought to φ with a belief that you have conclusive reason to φ. Second, it assimilates doxastic judgment to a kind of doxastic commitment, equating the doxastic judgment that p with the belief that there is conclusive reason or evidence for *p*.[131] I believe both assumptions are mistaken. First, though practical judgment includes a belief about your reasons, practical commitment is not commitment to it *qua* belief. Second, doxastic judgment – what you commit to when you form a belief – is not *about* your epistemic reasons but *informed* by them.

---

[129] Again, I claim only that abandoning a judgment *can* involve this species of self-mistrust. Of course, you can also abandon a judgment in the manner that Broome and Kolodny emphasize, by encountering a reason to think it mistaken without any self-mistrust. (Note that you can also follow through on a judgment because you've discovered a reason to think it correct. That is not to trust *that* judgment.) The proposal is that registering a reason to think a judgment mistaken is not the only way to abandon the judgment. (Likewise, registering a reason to think a judgment correct is not the only way to follow through on the judgment.) You might satisfy the requirement by self-mistrustful reasoning that takes you, in Kolodny's term, 'upstream.' (For an important caveat on the stream metaphor, see again note 2 above.)

[130] In *What We Owe to Each Other* (Cambridge: Harvard University Press, 1998, 25-30), Scanlon argues that we should restrict the term 'irrational' to cases in which your attitudes fail to conform to your judgments about reasons, but later in "Structural Irrationality" (in G. Brennan, R. Goodin, F. Jackson, and M. Smith (eds), *Common Minds: Themes from the Philosophy of Philip Pettit* (Oxford: Oxford University Press, 2007), 84-103), he merely distinguishes such *structural* irrationality from the *substantive* species of irrationality that I'm calling 'deliberative.' Since Scanlon agrees with Broome and Kolodny that only substantive (ir)rationality traffics in reasons, one could express my conclusion as the thesis that, properly understood, even structural (ir)rationality is substantive. But that formulation would be misleading. My principal aim is not to undermine the distinction between structural and substantive species of rationality but to recast it in a way that gives the concept of reasoning crucial work to do on both sides of the distinction. (Note that although Kolodny, citing Scanlon ("Why Be Rational?," 560, n. 49), speaks of a 'belief' about evidence or reasons (e.g. at ibid., 521), rather than (with Scanlon) of a 'judgment,' that merely reflects the fact that Scanlon does not himself draw the sharp distinction between judgment and belief on which I'm going to insist.)

[131] Broome's entire discussion in "Does rationality consist in responding correctly to reasons?," assumes that a plausible 'enkratic condition' would link commitment (whether doxastic or practical) to a belief about your reasons. And Kolodny's discussion of the 'core requirements' in "Why Be Rational?," section 5, explicitly embraces both assumptions.

Before pursuing these two objections, let's consider in more detail how Scanlon's equations fit together. On Scanlon's account of it, the rational force of a practical judgment figures as a species of doxastic commitment: by the first equation, your all-things-considered practical judgment that you ought to φ is the belief that you have conclusive reason to φ, which in turn, by the second equation, manifests commitment to the doxastic judgment that you have conclusive reason to φ.[132] If we accept Scanlon's equations we must notionally distinguish three things: (a) the doxastic judgment that you have conclusive reason to φ, (b) the doxastic commitment to that judgment that takes the form of a belief that you have conclusive reason to φ, and (c) the all-things-considered practical judgment that you ought to φ. If we equate (c) with (b), we have to remember that the only commitment in (b) is to (a). As far as the account goes, insofar as practical judgment embodies a commitment, it is not a practical commitment. Even if you are committed to (a), the doxastic judgment, by virtue of (b), the doxastic commitment, you are not yet committed to (c), the distinctively practical judgment. If we equate (c) with (b), a doxastic commitment, we should be explicit that neither of these amounts to a practical commitment – that is, to a choice or intention. We might try to view practical commitment as commitment to the doxastic commitment, but, as we'll see, that approach runs into problems.

We can begin to see the problems by considering how Scanlon's blurring of the distinction between practical and doxastic commitment figures in his argument against the possibility of upstream reasoning. Though Scanlon concedes that you may go irrational in the gap between practical judgment and intention, his view of how you are committed to a practical judgment ensures the illegitimacy of reasoning upstream. Here's how he puts it:

> [I]t would be irrational for an agent to avoid the incompatibility between judging herself to have compelling reason to do *A* at *t* and her not deciding to do this by abandoning the former judgment unless she saw some reason to revise this assessment. And it is difficult to imagine a case in which she could take her failure to decide to do *A* at *t* as a consideration bearing on the merits of doing it.[133]

Note how Scanlon treats the doxastic commitment at the core of a practical judgment – in our taxonomy, equating (c) with (b) – as committing you to a downstream-looking stance toward the question of practical commitment. Even if the practical judgment does not on its own commit you practically, he argues, it determines how you can reason through the question of practical commitment. You cannot reason upstream, the argument goes, because your practical judgment commits you doxastically, and this doxastic commitment cannot be undone unless you 'see some reason' – that is, some downstream-looking deliberative reason – to revise it. That in turn commits you to reasoning only downstream about your practical commitment.

Why cannot Scanlon's agent simply mistrust her judgment? Why must the mistrust involve a downstream assessment of her reasons? Scanlon seems to share the worry with which we began this section: such self-mistrust must register as evidence about your reasons, about which you can reason only downstream. Assuming, with Scanlon, that your practical judgment that you ought to φ manifests doxastic commitment to the judgment that you have conclusive reason to φ, it follows that a failure to commit to – that is, choose or intend in accordance with – your practical judgment amounts to a failure

---

[132] In "Structural Irrationality," Scanlon distinguishes "attitude-directed" from "content-directed" judgments (90-1), but the latter are nonetheless about reasons; what distinguishes them is merely that they are not *explicitly* about reasons *for other judgments*. He makes the distinction in reply to a worry that attitude-directed judgments are "somewhat artificial" (91): it's more natural to "direct our attention to the world" when we deliberate rather than to our further judgments. That's importantly true. But for our purposes it won't matter on which side of this distinction a given judgment lies.

[133] "Structural Irrationality," 96. Scanlon uses 'decision' to mark a practical commitment.

to commit *practically* where you are nonetheless committed doxastically. If, in accordance with the worry, you must view this failure of practical commitment as giving you evidence about your reasons, then you must count as reopening the doxastic deliberation informing your doxastic judgment that you have conclusive reason to φ. That amounts to a downstream assessment of your reasons.

We can use Scanlon's own framework to explain why the worry is confused. In the *status quo ante* you are doxastically committed to this judgment about your reasons: you do believe that you have conclusive reason to φ. You could, of course, undo that commitment through self-mistrust: mistrust now in the doxastic judgment that you have conclusive reason to φ. (I'll discuss that different case presently.) But it simply does not make sense to say that you 'mistrust' a belief.[134] What we're trying to say is that you mistrust the practical bearing of your belief: you mistrust it not *qua* belief but *qua* practical judgment. We're trying to say that you mistrust your practical judgment without mistrusting your doxastic judgment. But it follows from the fact that you do not mistrust your doxastic judgment that you *do* believe that you have conclusive reason to φ. The formulations implicitly concede that you do not regard the self-mistrust as giving you evidence about your reasons.

What the practical mistrust undoes is not that belief but its bearing on your choice or intention. It may take a further step, and perhaps a further degree of self-mistrust, now doxastic rather than purely practical, for you to abandon the doxastic judgment that informs that belief.[135] Until you abandon that judgment, you may be stuck judging that you have conclusive reason to φ while feeling no rational pressure to choose or intend to φ. And if you don't mistrust this doxastic judgment, you may even be stuck *believing* that you have conclusive reason to φ while feeling no rational pressure to choose or intend to φ.[136] But that's just to say that a practical judgment is not *simply* a belief about your reasons. A practical judgment may include a belief about your reasons, but it gives that belief a further practical orientation. The practical orientation is what brings you under **Wide Process**. What shows that there is this further practical question, beyond any mere belief about your reasons, is that you can persist in your best practical judgment that you ought to φ – that is, continue to believe that you have conclusive reason to φ – while nonetheless mistrusting that practical judgment. Since you don't thereby mistrust your belief – again, that simply doesn't make sense – you must be mistrusting an element in the practical judgment that is not included in the belief.

---

[134] As Wittgenstein epigrammatically observes, "One can mistrust one's own senses, but not one's own belief" (*Philosophical Investigations*, trans. G. E. M. Anscombe (Oxford: Blackwell, 1956), p. 190). And as Richard Moran observes, commenting on this passage, "this must mean not that I take my beliefs to be so much more trustworthy than my senses, but that neither trust nor mistrust has any application here" (*Authority and Estrangement* (Princeton: Princeton University Press, 2001), 75).

[135] How might a mistrustful reluctance to commit lead you, through upstream reasoning, to abandon your doxastic judgment? Though I lack space to pursue the issue, everything I've said about the self-trust dynamic on the practical side seems just as importantly true of the doxastic side. Just as judging that you ought to φ does not ensure that you choose or intend to φ, so judging that p does not ensure that you believe that p. If your self-mistrust gets the better of you, you thereby count as reopening deliberation, which entails that you've abandoned your judgment that p. But here again: that counts as upstream reasoning. Since judgment-to-commitment requirements have wide scope even in the doxastic case, you can satisfy them by reasoning your way from a failure to commit to your judgment to its abandonment. We can accurately describe the self-mistrust that gets the better of you as a failure to believe that p from which you reason back upstream.

[136] What should we say about the odd predicament in which you believe that you have conclusive reason to φ without that practical element, so you fall short of judging, all things considered, that you ought to φ? Well, the belief entails that you ought to make that practical judgment. But the belief is not itself a judgment that you ought to make that practical judgment. The belief is second-order, and it concerns your grounds for making the judgment. You might in the same spirit believe that you have conclusive epistemic reasons for the belief that p, yet fail to form the latter belief. There's an inconsistency here, but it doesn't involve the violation of the rational requirements that we're discussing in this paper. The requirements of rationality that we're discussing address the gaps that emerge within first-order thinking between judgment and commitment and further follow-through. We aren't discussing the requirements to keep your lower-order attitudes in line with higher-order attitudes about your grounds.

It is easy to misinterpret these self-trust and self-mistrust relations, whether doxastic or practical, if one follows Scanlon and equates judgment with a belief that you have reasons. We can see the problem most clearly by elaborating what I listed earlier as a second objection to Scanlon's account of judgment. If your doxastic judgment is the belief that there is conclusive reason or evidence for *p*, and your doxastic commitment is the belief that p, then your doxastic commitment has a different content from the judgment informing it. It thus becomes difficult to understand how your commitment could be a commitment to what you've judged. In the practical case, the parallel discrepancy is not so obviously a problem. Because we express the content of an intention with a verb alone – you intend *to* φ – it is not so mysterious how in forming the intention you commit to your practical judgment, whether we formulate the content of the latter in terms of a simple 'ought' or more elaborately with Scanlon. But if one starts from the doxastic, analyzing each with Scanlon, it becomes mysterious how you could commit to a judgment in either case, and one may simply lose sight of the self-trust dynamic. Of course, it does make sense to say that you commit abstractly to *p* or to φing, but the point of saying that is merely to individuate the commitment by specifying its content. This does not say what you commit to in terms of the self-trust dynamic; this is not the commitment that could lapse in akrasia. When commitment is considered as the counter to akrasia, what you commit to is what you could have mistrusted: not the content of the commitment but the judgment that informs it.[137]

[137] I suspect that the assumptions about judgment that I'm rejecting express a more fundamental 'internalist' thesis about the authority of judgment over the attitudes that it governs: that a judgment must somehow already 'include' the commitments that it governs, lest we be forced to posit an 'externally' mediating act or attitude whereby the judgment is made practically or doxastically effective. In "Receptivity and the Will" (*Noûs* 43:3 (2009)), I argue that the internalism in question can be vindicated without the assumptions. The bare claim that rational requirements are wide process requirements does not presuppose internalism, but it is important to see that it need not involve a conception of choice or intention as a faculty of willing. The wide-scope interpretation of enkrateia presupposes merely that the perspective of judgment does not simply determine the perspective of choice or intention (or, in a doxastic case, belief) in the dynamic of self-trust that it initiates. While akrasia is always as such irrational, the problem that it poses for the agent admits of more than one genuine solution.

Jorah Dannenberg
Stanford University

I am currently an Andrew W. Mellon Fellow in the Humanities at Stanford University. Prior to the fellowship at Stanford, I earned my degree from UCLA, where I wrote a dissertation on promising. My other philosophical interests include moral luck, the nature of the self and its role in ethical thought, and the history of ethics.

"Promising, Practices, and Interpersonal Obligation"

Abstract: *Social practice-based accounts of promising explain the obligation of a promise by appeal to facts about the entire community in which it is made. Scanlon and others have suggested that such accounts cannot explain the distinctively interpersonal character of the way a promise binds. I argue that this objection is rooted in a conception of the rules of our practice of promising that is unduly impoverished: it assumes our practice's rules must be specifiable in entirely non-moral terms. But a richer conception that makes use of moral terms in understanding the rules by which practitioners relate to one another was central to the social-practice view from its inception – it is suggested by Hume's account of fidelity as an artificial virtue. I argue that this richer conception better characterizes our actual practice of promising, and shows how that practice leads to distinctively interpersonal obligations between promisors and their promisees.*

Introduction

An important tradition in moral philosophy explains the import of a promise by appeal to facts about the attitudes and behavior of the entire community in which it is made: the so-called *social practice* of promising. A charge levied by those who favor rival accounts of promising is that practice-based views face some considerable difficulty in accounting for the way that breaking a promise represents a *distinct* wrong done to one's promisee.[138] The aim of this paper is, in one sense at least, rather modest: I want to show how a social-practice based account might successfully respond to this objection from the literature. What I offer thus falls far short of a full-fledged defense; the social practice explanation may succumb to any number of other problems in the attempt to provide the complete picture of how a promise binds. In another sense, however, my aim is ambitious. As I hope to show, the illusory force of this objection may stem from a rather deep distortion in how we have come to think of our practice of promising, and I suspect this distortion may carry over to other morally important social practices too. Whatever one thinks about the ultimate explanation of promising, much of moral life is undeniably a matter of engaging in social practices with one another. So the point I wish to make about the practice of promising is, I hope, of some interest beyond the debate about how we explain the duty to keep a promise.

Social-Practice Based Accounts of Promising

Hume is credited as the founding father of the social practice view. In the *Treatise*, he famously separated morality into its natural and artificial parts, and claimed that promising belonged on the artificial side of the divide.[139] The details of Hume's position are complex, but near the heart of his account is the following idea. Understanding the place of *fidelity* in the catalogue of human virtues requires recognizing the important social role of the distinctive motive we have to keep our promises;

[138] The objection is raised by Scanlon (1998) p. 316, and appears in Kolodny and Wallace (2003) pp. 125-6 as well as Tognazinni (2007) p. 203.
[139] Hume (1978). The discussion of promising occurs primarily in Book III, Part II, Section V.

in particular, the role that the prevalence of such a motive plays in making the members of a community fit for a life of ongoing social cooperation. Crucially, this role is not necessarily apparent in any single virtuous act of promise-keeping. One must zoom out and consider how things change when all (or at least most) of the members of the community come to have the motive as a standard part of their repertoire, in order to appreciate why promise-keeping matters to us in the way that it does.

In the 20th century, Rawls revived a version of Hume's idea in *Two Concepts of Rules*.[140] Rawls argued that the notion of a community wide practice of promising could be used by Utilitarians to fend off a standard objection: that the principle of utility could not account for the apparent stringency of the obligation to keep a promise. Breaking a given promise would often produce the best overall outcome, seemingly forcing Utilitarians to permit a class of actions that common-sense morality condemns. But Rawls pointed out that this objection elided the difference between giving a Utilitarian justification for the *practice* of promising as a whole, and justifying the claim that one must keep a particular promise by appealing to the rules of the practice. Properly understood, Rawls argued, facts about what would produce the best consequences were not among the considerations any given person participating in the practice of promising could consider in deciding whether or not to keep her word.

Later, in *A Theory of Justice*, Rawls presented another variant of the practice-based view.[141] Those who make promises, Rawls argued, voluntarily avail themselves of the benefits made possible by the ongoing social practice in their community. Promisors elicit the trust and cooperation of others in a way that would be impossible, were it not for the fact that the community as a whole engages in the practice. Given that this is so, Rawls argued it would be unfair for any person to make a promise, then fail to do her part in supporting and sustaining the practice by following its rules. Breaking a promise would amount to unfair "free-riding" on the willing cooperation of other practitioners.

These and other versions of the practice-view[142] differ in important ways, but they all share at least this common element: when I promise to water my neighbor's plants while she is on vacation, my obligation to keep my word cannot be explained merely by appeal to facts about just the two of us. Rather, the fact that she and I are involved in a community-wide, ongoing pattern of activity figures in the explanation in at least two ways. The first step in explaining why I am bound is to recognize that this is simply what, as "defined" by our shared practice, is involved when one has made a promise. That is, the immediate explanation of why I am obligated is simply that in our community, we regard a person who gives her word as bound to keep it.

Of course we need not end our inquiry with that answer: we may go on to ask a different kind of question, viz. why should *that* be our practice? What important social end does it further? Why (if at all) might it matter morally that anyone in particular does as our practice dictates? In order to answer this latter kind of question we can appeal to the sorts of general facts about the value of the practice as a whole: its role in fitting us for social-cooperative life, and/or to the general considerations about fairness that require that any one of us follow our practice's rules.

It is worth stressing that this two-stage[143] explanation seems to capture something deeply right. In the ordinary course of things, honoring our promises is often a matter of doing what we all do, because it's what we all do; at the same time, it is a deep and important fact about us that we all do it – that we are group of people for whom regarding our promises as important is second-nature. The fact

---

[140] Rawls (1999).
[141] Rawls (1971). The account is presented on pp. 344-348.
[142] I choose to highlight these three versions of the view in particular because of their prominence, and because I think they serve well to illustrate the central features of this type of position.
[143] The method of explicating the essential form of the practice account by distinguishing its distinct "two stages" is in Scanlon (1998), p. 295.

that as members of a social group each of us tends to be moved unreflectively to keep her word is a fact about ourselves that we can affirm and approve of when we reflect.

The Objection

Recently it has been urged that accounts of this general shape encounter considerable difficulty explaining how a broken promise represents a *distinct* wrong done to one's promisee. The thinking behind the objection seems to be this. The ultimate explanation of why keeping my promise to my neighbor is important will, at the end of the day, lead us to considerations about fair participation and/or facts about the value of the practice as a whole. But if these are the basic moral reasons that favor keeping my promise, then why should breaking it be a special affront to my neighbor in particular? These considerations of value and fairness that explain why I am required to do as our practice dictates would seem to leave every member of our community with the same cause to complain when I fail to do so. The basic moral considerations do not seem to concern anyone in particular. The practice provides a good for all; fair participation is owed to each.

This criticism may be thought especially forceful in light of the methodology implicit in much recent theorizing about promising, which assumes that explaining the moral significance of a promise is just equivalent to explaining how somebody does wrong by not performing the promised act. It would follow that what practice-based explanations of promising threaten to miss is the interpersonal character of promissory obligation as such - the fact that in promising one morally binds oneself *to* another person in particular, and not also in the same way to one's friends, co-workers, and anyone else who has ever made a promise in one's community. The distinctively interpersonal character of the bond that results from a promise seems essential; unless practice-based accounts can capture it, some other explanation of promissory obligation must be correct.

Seeing that the Objection is Mistaken

As formulated, the objection depends on a perfectly general feature of the practice-based account. I think that it is thus relatively easy to see something must be amiss. One need only call attention to other occasions where we undeniably *are* participating in some social practice, our reasons for adhering to that practice are likewise general, and yet we nevertheless recognize obligations of a distinctively interpersonal character arising as a part of what we do. So, for example, we raise our hands in order to be called upon to speak in a classroom, meeting, or other public space. If I interrupt you during your turn to speak, this can be a way of wronging you in particular. In certain contexts at least, such an interruption manifests not just a flouting of our practice that anyone present could object to (though it manifests that as well); my interruption shows a distinct lack of respect for you and your right to speak. Yet who would deny that hand-raising is a social practice, upon which my obligation not to interrupt you essentially depends? The practice facilitates orderly communication; it is also a good way to ensure that speaking time is equitably allotted. Facts like these explain why it's good to have the practice, and why any of us should do as it requires. Like the considerations thought to support the practice of promising, they appear to concern all members of the community equally. Evidently, the practice of hand-raising can nevertheless generate a distinctively interpersonal obligation, which I violate when I interrupt you during your turn.

Considering such examples suggests straightforwardly *that* this recent objection against practice-based accounts of promising must somehow be mistaken. Understanding *why*, however, is more difficult, and will be the task of the remainder of this paper. I think we can identify a particular assumption, upon which the objection depends for its illusory force. Rejecting the assumption will lead to a solution to the problem, by pointing us towards an account of our practice that makes it evident how individual promisors and promisees indeed come to stand in a distinctive kind of moral relationship with one another.

The Mistake and How to Correct It

I believe this objection is likely rooted in a conception of how we are to specify the "rules" of our moral practice of promising. In particular, the objection appears plausible because of a tendency to assume that in describing the rules of our practice of promising, we can or should somehow be able to do so in entirely non-moral terms. Construed this way, the *moral* part of the story is thought to come only *after* the practice has been completely described, when we turn our attention to the project of explaining how the practice as a whole is valuable, or how anyone might be morally obligated to follow its rules. But this conception is needlessly impoverished, and thus distorts our understanding of our practice. In contrast, as I'll go on to suggest, I think we can and should embrace a conception of our practice of promising in which the rules of our practice mandate how individuals are to treat and regard one another in irreducibly moral ways.

Games and sports are often among the first examples cited when explicating the concept of a social practice. In some ways, they seem well suited to illuminate many of the key ideas. Games and sports often involve highly specific and well codified rules, defining what it is to engage in them at all by, to paraphrase Rawls, giving structure to the activity through the definition of moves, roles, positions, penalties, defenses, and so on.[144] Moreover, games and sports can help to see the very stark distinction between asking a question like "what might be the point of us engaging in this activity?" and, on the other hand, "what is a particular player to do in a given set of circumstances?" The fact that the answer to the first question may be something like "the point of playing golf is to have fun" surely does not mean that on the golf course one determines what one is to do by asking "what would it be most fun to do now?"

On the other hand, perhaps the fact that our go-to examples of social practices tend to be games and sports can engender or reinforce the illusion that the rules of *any* practice, whether a game like chess, a sport like baseball, or a moral activity like promising, must always be describable in an entirely non-moral vocabulary. After all, the rules of games and sports invariably are. That is, the paradigm of a rule of a practice might seem to be something like the rule of baseball that defines what it is for a batter to draw a walk, a rule we might represent as:

After four pitches outside the strike-zone, the batter shall advance to first base

This rule defines what it is to draw a walk, by stating what someone playing the game of baseball is to do in a given situation. It invokes other terms and concepts that are also defined within the practice of baseball, like *strike-zone, batter, first base,* and even *advance* (in the unique sense in which one advances around the bases of a baseball diamond).

The focus on such examples might suggest that we should be able to describe what happens within any practice, and so in particular the practice of promising, in a way that is likewise morally unsaturated. Of course promising has no official rule book. But the practice view suggests that we should think of the activity as involving certain uncodified rules or norms, which characterize how the members of our community think we ought to act as promisors and promisees. At first blush, it may appear relatively straightforward that these rules can be represented as simply stating what a person is to do in a given situation, in much the way the 'walk' rule does. So, if we wanted to state the rule of our practice that defines what is involved in *keeping* a promise, we might imagine that something like the following does the job:

The promisor shall perform the promised act (unless the promisee waives)

[144] Rawls (1999) writes in the first footnote "I use the word " practice" throughout as a sort of technical term meaning any form of activity specified by a system of rules which defines offices, roles, moves, penalties, defenses, and so on, and which gives the activity its structure. As examples one may think of games and rituals, trials and parliaments."

Such a rule appears to say that keeping a promise is, more or less, doing whatever it was that was specified when making the promise, e.g. "watering the plants." It makes use of other notions internal to the practice of promising, like the roles of *promisor* and *promisee*, and the move *waive*. What it appears to provide is a description of what counts in our practice as "keeping" a promise, without any moral terms. That promising is a *moral* practice is determined only by what happens after this point, in thinking about what makes the practice as a whole important, or what gives us reasons to follow its rules.[145] Baseball is an engaging pastime; promising serves considerably more important social ends. A person who violates the rules of baseball is a bad sport; a person who violates the rules of promising does moral wrong.

Here is the problem: the suggested rule in fact does not do a very good job of describing what, in our practice, actually counts as *keeping* a promise. In particular, if our aim is to try to state what counts as a kept promise in the form of a rule, then I'm inclined to see something like the following alternative as getting us considerably closer:

> *Having made a promise, the promisor shall regard the promisee as having a certain status, such that the promisee is entitled to the performance of the promised act by the promisor; the promisee may choose to relinquish this entitlement.*

Let me first suggest something about the kind of reason there is for thinking that this second rule does better at capturing that part of our practice that concerns what counts for us as keeping a promise. I take this to be more or less an empirical claim, about how to describe the prevailing attitudes in our community about what makes for a kept promise. A recognition of a promisee's status as in some way entitled seems at the heart of what we actually think and do in our practice.

For example, our practice is not to count it as a *kept* promise if a promisor does what she promised to do only by coincidence. Suppose I promise to stop by your office to chat tomorrow at four o'clock. But then, completely disregarding my promise to you, I spend the afternoon wandering around campus aimlessly. Even if I happen to find myself at your door at just four o'clock, I have not kept my word. This is true, even though one complaint we often associate with the failure to keep one's word has no purchase here: the complaint that one was counted on to do something, and did not do it.

To point out another kind of example, our practice is not necessarily to regard a person as "off the hook" just because circumstances change in unforeseen ways that might make performance of the promised act no longer possible. If I promise to take you out to dinner at your favorite restaurant on your birthday in order to celebrate, I do not become free of my commitment if your favorite restaurant burns down, or you get the flu that day. I may keep the promise by taking you to a nice meal somewhere else that you like, once you're feeling better. That is, according to our practice, keeping one's word can, and often does, involve recognizing other ways in which one can make good, in light of changed circumstances.

I'm certain an even better formulation of the promise-keeping "rule" (or, more likely, rules) of our practice could be given. My point is simply that the second formulation above does considerably better than the first, by capturing that in our practice to keep a promise is not merely to perform the act described when one promised. Rather, it requires a certain kind of recognition for one's promisee as having a particular status – at a minimum, to keep a promise is to recognize that one ought to perform the promised act *because* one has promised it to one's promisee, thereby entitling her to its being done. In many cases, our practice seems to regard keeping one's word as involving considerably more. It

[145] That Scanlon (1998) conceives of the practice-based view in just this way seems evident in his discussion; the moral part comes only in the "second stage" of explanation.

seems to me that the second rule begins to represent this salient feature of our practice, while the first rule simply cannot.

The second rule can do so precisely because, in stating what it is to keep a promise, it deploys moral notions like *status* and *entitlement*; notions for which I think there is no evident kind of reduction to any more basic, non-moral description of the attitudes or behaviors required of promisors. Keeping a promise is, in this sense, not like drawing a walk in baseball –we simply cannot understand what, according to our practice, is involved in keeping a promise in fundamentally non-moral terms.

These moral concepts like *status* and *entitlement* have a place outside of our practice of promising: we recognize other situations in which there arise relationships in which we regard one another as having status, and there are other ways that one person can come to be entitled to the performance of some action by another. What it seems we should say is that our practice of promising takes these moral notions and puts them to novel use, by defining new and distinct moral relations between two individuals. The reasons to recognize and act on these new moral relations may stem from the general considerations about value and fairness that support the practice as a whole. But recognizing and acting on them for the relations that they are means seeing promisors as distinctively related to their promisees in irreducibly moral ways. In this way, the practice view captures how the obligation to keep a promise is distinctively interpersonal, even as the reasons that support the practice as a whole are not.

There is reason to think that this idea was near the heart of the practice view from its inception; recently we have lost sight of it. Something like it plays a central role in Hume's discussion in the *Treatise,* of how people might have managed to progress beyond the avowedly fictional "state of nature." Hume imagines how motives of selfishness and confined generosity would operate in conjunction with the scarcity of goods and their easy transfer without loss of value. Together, these psychological and material facts would have been a considerable impediment to forming larger groups, and reaping the many benefits of social cooperation. The denizens of Hume's state of nature solve this problem by forming rules that draw inspiration from the moral raw materials they already have, for example the natural bonds of family life. The result is new forms of moral interaction, built out of old: first property, then transfer by consent, then promise.

Emphasis on this aspect of Hume's account leads to a particular way of construing the role of interpersonal trust in the Humean account of promising. Hume's famous case in Book III, Part II, Section V of the Treatise involves two mutually disinterested farmers who face a coordination problem. Each would benefit if they exchanged assistance, but someone will have to go first. As Hume imagines it, absent a social practice of promising, whoever goes first will not have a good reason to expect that the other farmer will reciprocate, and so cooperation won't happen. The establishment of the practice of promising address the problem: the second farmer will have reason to reciprocate, since failure will result in the social penalties that attach to flouting the practice's rules.

The tendency, at least in recent discussions of promising, is to read Hume as though he regards these farmers as paradigmatic. When interacting with strangers, we face essentially the same predicament as Hume's farmers; luckily, we have an up and running practice, the role of which is to give potential promisees a reason to form the belief that promisors will perform the promised act when they otherwise would not have one. But read in light of Hume's discussion of the other artificial virtues, it makes more sense to think of the farmers as a representing a kind of prototype. Their problem is analogous to the strictly imaginary problem those in the state of nature face before they invent the institution of property. It is because we have a practice of promising that in our community otherwise mutually disinterested strangers are ordinarily able to elicit and receive trust from each other (where trusting a stranger cannot merely be reduced to simply forming a justified belief that she will do something).

The point deserves more attention than I can give it here: I raise it simply to provoke a different way of thinking about the Humean account. The crucial point for my purposes is that, read this way,

the bit of social artifice that creates the practice of promising surely does not *invent* the notion of interpersonal trust, but rather co-opts it (in a non-pejorative sense). Interpersonal trust already exists within the family, or among friends. What the artifice of promising creates is a new way of making *use* of trust, thereby enabling, for example, mutually disinterested strangers to enter into the kinds of cooperative projects for which trust is a precondition. It does this by establishing rules for when trust can be formally elicited and accepted. People involved in the practice now have a new way of coming to trust one another.

There may be much in the details of Hume's "just-so" stories of moral artifice that we may not want to embrace. His problematic is bound up with his sentimentalist account of morality, and his impoverished picture of human motivational psychology. But it seems to me that this idea is one those who would make social practices an important part of the explanation of moral life should want to resuscitate. The *raison d'etre* of invoking a social practice in order to understand some morally important activity need not be showing how something moral can be built out of entirely non-moral parts, but rather how moral parts can be recombined to make something new.

Works Cited

Hume, David. 1978. *A Treatise of Human Nature,* edited by L. A. Selby Bigge and P. H Nidditch. 2nd ed. Oxford: Oxford University Press, 1978.

Kolodny, Niko, and R. Jay Wallace. 2003. Promises and Practices Revisited. *Philosophy and Public Affairs* 31 (2): 119-154.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

———. 1999. Two Concepts of Rules. In *Collected Papers*, 20-46. Cambridge: Harvard University Press.

Scanlon, T.M. 1998. *What We Owe to Each Other*. Cambridge: Harvard University Press.

Tognazzini, Neal. 2007. The Hybrid Nature of Promissory Obligation. *Philosophy and Public Affairs*. Vol. 35, No. 3: 203-232

Keynote Address:

T. M. Scanlon

Harvard University

T.M. Scanlon is Alford Professor of Natural Religion, Moral Philosophy, and Civil Polity. He received his B.A. from Princeton in 1962 and his Ph.D. from Harvard. In between, he studied for a year at Oxford as a Fulbright Fellow. He taught at Princeton from 1966 before coming to Harvard in 1984.

"Ideas of the Good in Moral and Political Philosophy"[146]

My topic is the relation between ideas of the good that a person should use in assessing his or her own life and ideas of the good that figure in moral and political philosophy.

Moral Philosophy, as I will understand it, is concerned with principles regulating our conduct toward one another. It is concerned both with the *content* of these principles—with what morality requires—and with their *ground*—with why we should care about what morality requires. Political philosophy, as I will understand it, is concerned with standards for assessing large scale social institutions that we participate in and expect others to participate in. Here again there are questions of content—what justice requires—and questions of ground—why we should care about justice.

In both cases, answers to questions of content and questions of ground seem to involve or depend on claims about what is good for individuals. But it is important to distinguish between to different kinds of claims about the good for an individual. Claims of the first kind are claims about what is good *from that person's point of view*, that is to say about what an individual has reason to want. Clams of the second kind are claims about what benefits a person, or makes his or her life go better. We may call claims of this kind claims about *a person's good.*

The normative standpoint defined by the question of what is good from an individual's point of view is in one way an objective standpoint since it takes into account reasons provided by considerations other than the ways that person's life may be affected, such as reasons provided by the good of others. But this point of view remains subjective in another sense since it is defined by the question of what *that person* has reason to want and to do, and the answer to this question will depend on that person's particular situation, aims, relationships and so on.

The content of moral principles and principles of justice clearly depends on facts about individual good. Justifications of particular principles of right and wrong as morally correct must appeal to the ways in which individuals' lives would be affected if people generally abide by those principles of if they are free not to. They appeal, for example, to the reasons individuals have to want not to be subject to threats and violence, not be able to rely on agreements they have made, and so on. Utilitarian theories justify principles by appeal to such effects, but so do non-utilitarian theories such as my contractualism, and Rawls' theory of justice, although they do this in different ways.

What alternative is there to this way of justifying conclusions about right and wrong? Appeal to moral intuition is one possibility. But it seems possible to give reasons for our intuitive judgments

[146] This paper was first presented as the Routledge Lecture at the University of Cambridge. I am grateful to members of the audience on that occasion for their comments, and to Derek Parfit, Frances Kamm, Christine Korsgaard and Daniel Star for helpful comments on a later version.

about moral right and wrong, and these reasons depend in large part to what it would be like to live under such principles, and what things would be like if accepted standards were different. What else could they appeal to? The will of God might be suggested as an alternative. But would obeying God's commands make sense if these commands were arbitrary? It is surely important that they are the commands of a *loving* God, and hence grounded in concern for us. So on this view as well, the justification for the content of moral principles will depend on the way they benefit us, and hence on some conception of our good.

Conceivably, social institutions and the content of moral requirements might be justified by appeal to some good other than the good of those to whom these provisions apply. Indeed, in the case of some institutions this seems the appropriate mode of justification. The content of my rights and duties as a Harvard faculty member, such as the right to speak and vote at departmental meetings, are justified not by *my* interests but by what is necessary in order for the University to operate in a way that serves its goals. These include promoting the good of other individuals by providing education, and useful research, but also producing advancements in knowledge and understanding that are good in themselves, apart any from any practical benefits they it might lead to.

In the case of large-scale social institutions such as the state, however, justifications of the latter sort, which appeal to impersonal values, have something of a bad name. "Perfectionist" is one thing that such views are sometimes called. In contrast to such views, Rawls says at the outset of *A Theory of Justice* that he is viewing a society as "a cooperative venture for mutual advantage." (*TJ*, 4-5.) This may seem an unexceptionable remark, and he offers no justification for it. But it is a substantive assumption, in favor of what might be called *individualist* answers at least to questions of content. My immediate response is to accept this assumption, and count myself an individualist in this sense—that is, as someone who believes that in both moral and political philosophy answers to questions of content must be based on claims about the good of individuals. The remainder of this paper is in part an inquiry into the degree to which this is so: the degree to which individual good is the fundamental notion for moral and political philosophy and, to the degree that it is, how the conception of good that plays this role is related to the conception that is relevant for an individual who is making decisions about his or her own life.

Consider now questions of ground—of what reason we have to accept moral principles as binding on us, or to care about principles of justice. Answers to these questions will need at least to take into account claims about individual good. They will need to explain how being just, or refraining from treating others in ways that are morally wrong, is at least compatible with having the kind of life that an individual has reason to want for him or herself. But answers to questions of ground cannot, I would say, appeal *only* to claims about individual good. Our reason for accepting moral principles as binding on us and for wanting to live under just institutions is not simply that these things contribute to our good. At least, insofar as these things do contribute to our good this is because they are ways of living that we have other reasons to want, reasons that are not entirely reducible to considerations of what benefits us. So answers to questions of ground will be, in the first instance, claims about what is good *from an individual's point of view*. They also *involve*, that is to say, at least take into account, claims of the narrower sort about individual good. But how claims of these two kinds are related in general, and how they interact in answering questions of ground, are complex questions.

Leaving these complexities aside for the moment, however, it at least seems clear that answering both questions of content and questions of ground involves making claims about individual good, even if these are not the only claims involved. What I am interested in, as I have said, is how the ideas of good that are relevant to these answers are related to each other and to conceptions of good that individuals have reason to employ in making decisions about their own lives. It might seem that there should be a single conception of good that can serve all three of these functions: as the basis of individual's own personal decisions about their lives, as a basis for argument about the content of moral principles and principles of justice, and as the basis for arguments about why individuals should

care about such principles. I will argue that this is not the case: the conception of individual good that figures in arguments "from the moral point of view" as it were, about the content of moral requirements differs from the conception that individuals rightly appeal to in guiding their own lives.

If, however, there is this difference between the conception of good relevant to answering questions of content and the conception relevant to an individual's decisions about his or her own life, this raises a question of ground: why should individuals take seriously a form of moral justification that does not take account of his or her own life in the way that is most directly relevant for them? This question arises in different forms for different theorists, depending in part on their conceptions of individual good and on their ways of understanding the relevant "moral point of view." I want now to consider how this question arises and is dealt with by a number of apparently diverse thinkers: Henry Sidgwick, Karl Marx, John Rawls and Bernard Williams. I will maintain that despite their diversity there is a striking similarity in the problems they face.

Consider first the way in which this question arises in Sidgwick's discussion at the end of *The Methods of Ethics*. Sidgwick distinguishes two rational points of view: two ways in which Reason can arrive at conclusions about what one ought to do. From one point of view we consider only how our own life is affected. From the other we take into account affects on the lives of all sentient beings. We may refer to the first of these as the point of view of one's own good (what is good *for one*, in the terms I used above), and refer to the latter as the point of view of what is good, generally. So described, it may seem as if conclusions reached from these points of view cannot conflict—they are conclusions about two different things: about what is good for me and what is impartially good. But as I have said, Sidgwick understands these as two ways in which Reason can arrive at conclusions about what one ought to do. Since what is best for me is not always the same as what would be impartially best, conclusions about what one ought to do reached from these two points of view can conflict. When they do, Reason gives us conflicting directives about what we should do: Reason is, as Sidgwick says, "divided against itself." (508) He saw this as a crisis, which he referred to as the Dualism of Practical Reason.

Although it is not essential to this crisis, which could arise under different understandings of what is good for me and what is impartially good, it will be helpful for what follows to say more about how Sidgwick understands these two notions. First, he believes that what is of ultimate value in both cases is agreeable consciousness: in the case of my good, my own agreeable consciousness and in the case of impartial good the agreeable consciousness of any sentient being. Each form of good is a mathematical whole composed of occurrences of agreeable consciousness, and in each case Reason requires a form of impartiality. In either case, it is irrational to prefer a unit of agreeable consciousness at one time to an otherwise identical unit occurring at some other time, and from the point of view of what is good, generally, it is irrational to prefer a unit of agreeable consciousness occurring in one life to an otherwise identical unit occurring in some other life.

Since what Utilitarian morality requires of us is that we strive to produce the greatest quantity of happiness for all sentient beings, what we ought to do from the impartial point of view coincides with Utilitarian Duty. So the conflict that troubles Sidgwick can be described in two ways: as a conflict between two dictates of Reason, and as a conflict between (Utilitarian) Duty and self-interest[147].

Although what is ultimately valued from these two points of view is in one sense the same—agreeable consciousness—it is important not to overlook the difference in the *way* it is valued. What is good *for me* is an additional unit of agreeable consciousness in *my* conscious life. What makes things better from the point of view of the universe is an additional unit of agreeable consciousness in *someone's* conscious experience. To put the matter in the terms used by Thomas Nagel in *The Possibility of Altruism*, reasons of self-interest are *subjective* reasons, requiring a reference to the person in question, whereas reasons to promote the good generally are what Nagel called *objective*

[147] As Sidgwick does, for example, on pp. 502-506 of *The Methods of Ethics*.

reasons: the facts that are such reasons can be described, in a way that captures their normative significance, without reference to the person for whom they are reasons.[148]

The conflict that troubled Sidgwick depends on the independence of these two forms of value. If a unit of agreeable consciousness were *good for* me only if, *and because*, it is good impartially (good from the point of view of the universe) and happens to occur within my life, then the conflict would not arise. I will call this the Moorean thesis, after G. E. Moore, who espoused something like this view.[149] If this thesis were correct, then conclusions about m own good would simply be partial conclusions about what is good impersonally (they would be conclusions about what is good in one part of the universe, so to speak.) Such conclusions would therefore have no normative weight against conflicting conclusions about what is impersonally good, on balance. They would simply be subsumed by judgments of the latter kind, and the problem Sidgwick faces would not arise.

I believe, as I have already indicated, that in many cases what is good *for* me depends on what is good in a way that does not depend on me. The degree to which success in my aims is good for me (makes my life better) can depend on the degree to which these aims are worth pursuing, apart from any benefit to me. Success in finding a cure for cancer, or proving a deep theorem would make my life better. Success in counting blades of grass would not.

It may be maintained that curing cancer and proving deep theorems are things worth doing only if, and because, they lead to results that are good for others (make their lives better.) On this view the idea of what is good for some individual is normatively basic. I am not certain whether this is true.[150] Even if it is, however, the present point remains: that the degree to which success in one's pursuits makes one's life better depends on whether these aims are worth pursuing for reasons that do not depend solely on effects on one's own life.

We can maintain that this is so while still holding that something's being good for a person does not, in every case, depend on its being impartially good, in the way that the Moorean thesis claims. This dependence seems particularly implausible in the case of the kind of good Sidgwick took to be basic: agreeable consciousness. A unit of my agreeable consciousness is not good *for me* only because, as a unit of *someone's* agreeable consciousness, it is good impersonally, and it happens to occur in my conscious life.

My claim that this is not so may or may not be compatible with what Thomas Nagel argued in *The Possibility of Altruism*. Nagel's claim was that we must be able to see what is good for us as also good impersonally, and hence to see, and to be motivated by the fact that, similar things in other people's lives are also good in the same way. So far, this is compatible with what I have just said, because it does not mean that we have to see something as good for us *only because* it is an occurrence in our lives of something that is impersonally good. But Nagel does also say that any subjective reason has to be statable in objective form *with the same motivational content.*[151] So the compatibility of his view with the claim I have just made is not entirely clear. (But it should also be said that this claim of Nagel's is one of the most controversial in his book.)

---

[148] Nagel, *The Possibility of Altruism*, pp. 90ff.

[149] This Moorean thesis is defended by Donald Regan in "Why Am I my Brother's Keeper?" in Wallace, Pettit, Scheffler and Smith, eds., *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, pp. 202-230. For criticism of Moore's claim that *good for* is unintelligible, see Richard Kraut, *Against Absolute Goodness* (Oxford: Oxford University Press, 2011), Chapter 12.

[150] For a thoughtful defense, see Richard Kraut, *What Is Good and Why* (Cambridge, MA: Harvard University Press, 2007), esp. Chapter 4 and pp. 268-269. In assessing this issue it is important to bear in mind the distinction made above between what is *good for* an individual (makes his or her life better) and *the good for* an individual (what he or she has reason to want.) It seems obvious that something is good only if it is part of the good for some individuals in the latter sense. The question is whether it must also be *good for* some individuals in the former sense.

[151] *The Possibility of Altruism*, pp. 110-115.

The crisis for Sidgwick represented by the Dualism of Practical Reason arises partly because he focuses only on two modes of practical thinking—asking what is good for oneself and asking what is good impersonally—and he recognizes no mode of practical reasoning within which conflicting conclusions of these two modes of thinking can be adjudicated. Things look quite different if we take the most authoritative mode of practical thinking to be that within which we ask what we have most reason to do all things considered. The point of view defined by this question takes into account, within one normative frame, so to speak, what is good for us (what we have reason to want because of how we are affected by it) and what is good for others (what we have reason to want because of how they are affected by it), along with any other reasons we may have. Conclusions reached by considering this question have rational authority, since it would seem that Reason requires us to be guided by conclusions of this kind (not to do so would be irrational.) Recognizing the rational authority of this wider point of view involves downgrading the authority of the two points of view that Sidgwick considered, since we can ask from the wider standpoint what kind of importance to give to their conclusions, and in particular what we have reason to do when these conclusions conflict.

If conclusions reached through Sidgwick's two modes of practical reasoning are understood as two kinds of conclusions about what is good, then the question that arises from this wider standpoint is what reason we have to pursue what is impersonally good in this way at a certain cost to what is good for us (or to do the reverse.) If, on the other hand, Sidgwick's impersonal point of view is seen as (one way of understanding) the moral point of view, then this question becomes a version of what I called above the question of the ground of morality.

This question—of the ground of morality—is a question about what is good *from an individual's point of view*, but an answer must, as I said earlier, take into account what is good *for* that individual—it must recognize how giving a certain priority to conclusions reached from the moral point of view affects the quality of a person's life, and must explain how, because of and in spite of these effects an individual has reason to give morality's demands this priority.

Such an explanation will depend on the way in which both the moral point of view and the good of an individual are understood. Moving away from agreeable consciousness as the basis of these points of view opens up a wider range of possible answers to these questions. But any plausible account of the moral point of view (any answer to the question of content) must include *some* way of understanding the good of each individual. So the question remains how this conception of individual good is related to an individual's conception of his or her own good. Insofar as these conceptions differ, this difference makes the question of ground more difficult to answer. I want now to consider this difficulty as it arises in political philosophy as a problem about religious toleration discussed by Karl Marx and by John Rawls.

It is generally agreed that justifiable political institutions must allow individuals to practice their various religions as long as this does not involve practices that interfere with the lives of others. As I would put it, a principle that counted institutions that did not do this as entirely just would be a principle that it would be reasonable to reject. This rejection would be reasonable because individuals have reason to want to be able to practice their own religion. But *religion* as it is employed in this reasoning, is not a category that is important in the personal thinking of a religious person. It applies to, and treats as having the same importance, a wide variety of different views, many of which are, from the point of view of an adherent of any one of these doctrines, false or even pernicious.

This point, about the contrast between the personal outlook of a religious person and the argument for religious toleration, was addressed by Marx in his famous 1843 essay, "On the Jewish Question."[152] Marx was responding to Bruno Bauer, who said that the idea of religious toleration (in the case at issue, treating Jews as having the same rights as any other citizens) was incoherent. This was, Bauer maintained, because religious toleration involved seeing various religions as having equal

[152] In Robert C. Tucker, ed., *The Marx-Engels Reader* (New York: W.W. Norton, 1972), pp. 24-51.

standing, and one could not take this view without ceasing to be religious. (In particular, he said, Jews who saw things in this way would ceasing to be Jewish, by regarding Judaism as merely one view among others.) Marx replied that Bauer failed to see that religious toleration involved only declaring the difference between different religions to be *politically* irrelevant. That is to say, irrelevant to a person's political standing. This did not, he said, involve declaring the difference between religions to be irrelevant any more than getting rid of the property qualification for voting would involve abolishing property.

One should, Marx said, distinguish two standpoints: the standpoint one takes in one's personal life, from which one is guided by the tenets of what one takes to be the truth about God and human life—by the tenets of Judaism, as it may be, or Roman Catholicism—, and the standpoint one takes as citizen, when one is applying the law or arguing about what the law ought to be, where one is guided by the idea that religions should equal before the law, and that no one should be denied rights because of his or her religious beliefs. One can hold the latter view for political purposes, Marx said, without ceasing to take one's own religion seriously in one's private life.

John Rawls distinguishes two similar standpoints. Each person, he says, has his or her own religion or, more generally his or her own comprehensive view about matters such as the meaning of human life and its place in the universe. But, although we are each guided in our private lives by our own comprehensive view, when we are considering what our basic social institutions should be like—when we are addressing what he called constitutional essentials and questions of basic justice—we should be guided by reasoning that appeals not to values peculiar to our own particular comprehensive view but rather to what he called political values, which all can recognize, regardless of their comprehensive views. The importance, for any person, of living in accord with his or her religion and, as Rawls puts it more generally, the higher order interest of each individual in developing and pursuing his or her own conception of the good are examples of such values.[153]

Like the two modes of reasoning that Sidgwick described, the modes of reasoning we engage in from these two standpoints—the standpoints of man and of citizen, as Marx calls them—begin with values that are both similar and different. They are similar in that they value the same things, such as an individual's being able to live in accord with the tenets of his or her religion. But they value these things in different ways. An individual values living in accord with the tenets of his or her religion, or the ideals of his or her comprehensive view, because that is the way he or she believes he or she has most reason to live. In justifying basic political institutions, on the other hand, we place importance on their allowing individuals to follow their own religions or their own comprehensive views *whatever these may be* because individuals have the strong reasons just mentioned for wanting to do this, because what they have these reasons to want to do are different, and because treating their diverse reasons as on a par (for purposes of political justification) is a way of treating all citizens as equals.

There are, potentially, two kinds of tension between these two normative standpoints. If our institutions allow others to live according to their own conceptions of the good, we are likely to find ourselves surrounded by people who live in ways that we disapprove of. If this happens, then we cannot have the kind of public space that we would most like to have, for ourselves and our children. Call this the *empirical* tension, because it depends on what actually happens. The second kind of tension, which I will call *normative*, consists in holding the view that the reasons provided by one's religion (or one's comprehensive view), which one regards as of fundamental importance in one's own life, do not apply in the justification of political institutions. This is a tension one may feel even if religious toleration turns out to involve no empirical cost, because one happens to live one's life in a society in which, at least for a time, almost everyone holds the same comprehensive view one holds oneself.

[153] Cite *Political Liberalism*.

Rawls believes that if our comprehensive view is what he calls *reasonable*—if it supports the right idea of respect for others, including others who do not accept it—then even if religious toleration turns out to involve empirical tension, normative tension will not arise, because our comprehensive view itself will dictate that we should limit the application of our other specific beliefs to basic political questions.

Rawls believes that this kind of reconciliation between the two standpoints is the best we can hope for. This is an essentially liberal position. Marx was not a liberal. Although, unlike Bauer, he recognized and valued religious toleration as the best we can do under current conditions, he did not see it as the best we should hope for. What we should hope for, he thought, is not mere political emancipation but what he called human emancipation, in which religion is eliminated (along with property and the state), thereby eliminating the alienation involved in the tension within us between the outlook of man and that of citizen.

A formally similar internal tension lies behind my second example, from moral philosophy. This is Bernard Williams' famous remark about "one thought too many." Williams was discussing the views of Charles Fried, who in his book, *Right and Wrong* considered a case in which a man in a burning building had to choose between saving his wife and saving a stranger. Fried was discussing this question within the context of what might, I suppose, be called a neo-Kantian moral theory. He said, plausibly (?) that the equality of persons that such a morality involved did not imply that it would be impermissible for the man to save his wife, rather than saving a stranger, or deciding between them by flipping a coin. The reason, Fried said, was that morality must allow people to give special concern to their loved ones. Williams' characteristically sharp retort was that in suggesting that the man's reasons for saving his wife was "because she is my wife" this theory would give the man "one thought too many." The man's thought, if he even has to think rather than just to act, should be, rather, "My God! It's Anne!" Or something like that.

Williams' objection seems to me to miss the mark in a way that is similar to Bruno Bauer's. In framing moral principles or legal policies, we need to take account of the fact that individuals have special reasons to be able to help, and to favor, their family members and others close to them. But this abstract categorization of those reasons—as reasons to help and favor "close friends and family members"—is a way of recognizing, and equating for moral purposes, the particular reasons of particular individuals to save particular people close to them. Like the category, *religion*, it is abstractly formulated precisely to play this equating role, for the purposes of moral justification. But this is not to suggest that any individual, performing an action recognized by this category, would be moved by this abstract reason. Its relevance is moral, not personal.

Williams is not blind to this distinction. He is arguing that we do not have good reason, from our personal points of view, to give the categories of moral thinking the weight normally claimed for them. In formulating his argument he seems to me to overstate the kind of significance that morality needs to claim for itself. But, like Marx in the later part of "On the Jewish Question," he might well hold that recognizing even the more limited significance that I have claimed for abstract moral categories involves having an undesirably divided self. The cost of accepting the constraints of morality does not lie only in the sacrifices it might require on particular occasions—on occasions when we are not permitted to save our spouse because the cost in other lives is too great. His idea seems to be that, whether or not such occasions ever arise, accepting this as a *possibility*–accepting morality as a limitation of our commitment to our loved ones and our "ground projects"—is alienating because the normative tension it involves blocks us from whole hearted versions of these commitments. This is a negative claim about the question I have called the ground of morality: a claim that we cannot answer that question positively about a morality that involves this kind of alienation from our personal relations and personal projects.

All three of the problems I have considered—Sidgwick's Dualism of Practical Reason, the problem of political justification discussed by Marx and Rawls, and Williams' challenge to individual

morality—posit two different standpoints between which conflicts can arise, and in the cases of Marx, Rawls and Williams, a third standpoint that dictates how this conflict should be resolved. I want to consider now the different ways in which these conflicting standpoints can be understood, and consider in more detail the nature of the possible conflicts between them.

Sidgwick characterizes the content of his two forms of reasoning in the simplest terms. What is good for a person is determined by the quantity of agreeable consciousness in his or her life, counting the same quantity of consciousness equally whenever it occurs. What is good impersonally (and the standard of duty) is the greatest amount of agreeable consciousness overall, counting equal quantities of consciousness the same, whatever life they occur in.

If we move beyond agreeable consciousness as a basis both of individual good and of what is good from the moral or impartial point of view, this opens up more complex relations between the two standpoints and new possibilities for reconciling them. Let me begin by sketching an alternative account of individual good.

I believe that any plausible account of the good for an individual will be pluralistic, including factors of the following three kinds.

First, *experiential* factors, such as enjoyments, excitement, and the absence of states such as pain and fear.

Second, valuable relationships, such as friendships, good relations with one's family, and with others, such as those with whom one engages in cooperative activities.

Third, success in one's main aims and projects, insofar as these are worth having.

I do not have a systematic argument that things of these three kinds are good for a person and make his or her life better.[154] But it seems to me clear that they do make a person's life more choice worthy, and that any account of what is good for a person that left them out would be flawed and incomplete. Moreover, it seems clear to me that these three forms of good are independent: none of them is fully reducible to any of the others: the goods of achievement and relationships are not, for example, explicable simply in terms of the experiences they involve or foster. Monistic experientialism does not seem to me at all plausible. Nor does it seem to me that there is likely to be a plausible monistic account of any other kind, that is to say, any other good not on this list, such that all of these things are good for a person only because they lead to or are required by this further good. As I said, I cannot prove that this is so, but it is how things seem to me. What I want to do now is to draw out from these assumptions some conclusions about the relation between what is good for and individual and what an individual has reason to want more generally.

As I said earlier, I reject the Moorean thesis, according to which something is good for a person only if and because it is something that is good impersonally (a good thing to occur in the world) and occurs in that person's life. But any conception of individual good that includes the three elements I have just listed will make what is good for a person depend, in more complex ways, on what that person has reasons to want or to do that are at least in part independent of that person's life. This is most obviously true of the third category. Success in a pursuit makes one's life better only if it was something worth pursuing. Otherwise it represents a wasted life. And in most cases the reasons that make a pursuit worthwhile will have to do with factors other than effects on the life of the person in question, such as the effects on the lives of others (as in the case of searching for a cure for cancer) or the excellence of the skill or knowledge that one is striving to develop (as in intellectual and artistic pursuits.)

---

[154] For discussion see Derek Parfit, *Reasons and Persons*, Appendix I; Joseph Raz, *The Morality of Freedom*, Chapter 12; T. M. Scanlon, *What We Owe to Each Other*, Chapter 3; Stephen Darwall, *Welfare and Rational Care*. Richard Kraut offers an account of individual good that is also pluralistic in its content, but unified within an overall idea of flourishing. See his *What Is Good and Why* (Cambridge, MA: Harvard University Press, 2007), Chapter 3.

This is true as well, although less obviously so, in the case of some goods in my second category, of relationships. A relationship makes a person's life better only if it is one worth having (as opposed, say, to relations of servility or the devotion of fans to movie stars.) It is plausible to say that in these cases what makes these relationships worth having, or not, is the simply effects they have on the life of a person who enters into them. But this does not seem to me always to be the case.

I believe that the best answer to the question of the ground of morality lies in this area. Individual morality, as I understand it is, most fundamentally, a matter of being concerned to act only in ways that are justifiable to others. This is something one has reason to do because of the kind of things other persons are. As creatures capable of understanding and responding to reasons, they are creatures we have reason to want to be related to on a basis that is justifiable to them. Since this is a valuable relationship, achieving in it also makes our lives better, even though our reasons for wanting to achieve it derive in part from the kind of things they are, not simply from considerations of our individual good.

The problems that I discussed arising for Marx, Rawls, and Williams had to do with the abstractness of the considerations that figure in determining the content of morality and justice, as compared with the more concrete and specific ideas of value that individuals properly hold. It is plausible to think that this abstractness arises from the specific demands of moral justification, and hence, as I said, to put added pressure on the question of ground—the justification for taking "the moral point of view" or the point of view of justice, seriously. I want now to consider the degree to which this is the case.

Sketch of individual good that I have given also deals with abstract categories: pursuits, relationships, etc. This is not only because it is just a sketch. Any general account of individual good will need to be filled in by "subjective factors," because there are many worthwhile pursuits and many relationships worth having. The quality of a given person's life will depend on success of his or her particular aims, on the quality is his or her actual relationships with particular people, and on what happens to these particular people. The appeal to abstract categories is not due to the demands of a "the moral point of view" that treats people equally. Even Williams' conception of individual good involves to such categories, such as the idea of individuals "ground projects."

Things will be different insofar as there are aims that everyone has conclusive reason to pursue, or relationships that everyone has conclusive reason to have. In such cases there will be no need, or room, for subjective factors—everyone's life is made better, or worse, by the same standard. Morality, on the account I have just sketched, will be an example of this: anyone's life is made worse by failure to take seriously moral obligations to others.

Any *general* account of "the good from an individual's point of view" (of what an individual has reason to want) will also be abstract in the way just described: an individual has reason to do what is needed to promote his or her main aims, insofar as these are rational, and to promote the good of those with whom he or she stands in certain special relations: friend, spouse, child, brother or sister and so on. This is as much as we can say from the point of view of a general account that applies to everyone. But from an individual's own point of view these categories are less important. Not that they are never relevant: sometimes one acts from reasons of obligation deriving from such a relationship: "I had to help him," you might say, "after all, he's my brother." But Williams' point is that, ideally, these are not one's only reasons. To love someone involves caring about *them* and seeing what happens to them as in itself providing one with reasons, not dependent on the objective frame of family or other obligations.

These objective factors become relevant again when we are justifying claims about the content of morality. The basic facts relevant to such justification are in the first instance facts about what individuals in certain circumstances have reason to want, and to do. But not just any reason is relevant to determining what is *owed to* a person. What is relevant to determining what we owe to each other are reasons grounded in the way we would be affected by given principles. Relations such as "friend,"

"spouse," or "family member specify ways of being affected. It is not that one's friends and family members count morally only insofar as they stand in this relation: they count morally in their own right, and this is specified by what is owed *to them*. But what is owed *to you* must take into account the special reason you have to want to be able to help them when the need arises, and this reason, insofar as it specifies how *you* are affected, must have to do not with the fact that it is *them* but with their relation to you. But, from the fact that such a relationship has this kind of significance in moral justification nothing follows about the reasons that move you to help the individuals to whom you stand in this relation. Ideally, these individuals will be especially important *to you*, and the fact that it is *they* who are affected, rather than a fact about their relation to you, will be the main thing that moves you to act. But your special concern for them need not involve holding that, morally speaking, more is owed to them than to others. (Nor need it involve a similar claim about the good: that it is a worse thing if something bad happens to them than if it happens to someone else. There is a difference between what one must, as a friend, hope and wish for, and what one can sensibly think of as better.)

If you do not hold that your friend is morally special in this way, then the moral importance that they have in their own right (equal to that of others) will capture all the remaining moral importance that you attribute to them beyond the special importance that they have in determining what is owed to *you*. Given these assumptions, Williams' worry is not troubling.

But these assumptions are not trivial. A person might hold not merely that he has strong reason to be allowed to help his friend rather than a stranger, but that his friend is uniquely morally important: what happens to him or her matters more, morally, than what happens to other people, and that the reasons that support giving him or her this importance apply to everyone. In this case there will be an incompatibility of the kind Williams describes between your concern for your friend and a fundamentally egalitarian moral view.

What we might say (presupposing such a moral view) is that this problem will not arise as long as a person is reasonable (as measured by such a view) in what she claims for her friends. Alternatively, we might say that we need not worry about such conflicts because a relationship with a person that involves attributing this kind of unique moral importance to a him or her is not a relationship anyone has good reason to have: the relationships that can form part of an person's good are constrained by their internal compatibility with (egalitarian) morality.

The problem that Marx and Rawls discuss has a similar structure, up to a point. In that case, we start with the idea that the fact that something is *my religion* gives me a reason to object to a principle or policy that would prevent me from practicing it. So far, this is parallel to my reason for objecting to a principle that would prevent me from helping *my* wife rather than a stranger. As in that case, this does not mean that *my* reason for wanting to live up to the requirements of, say, Roman Catholicism, is that it is *my religion*. I may have strong positive reasons for accepting and valuing it, rather than some other faith (just as I value my friends, themselves, and not only under the relational category *my* friend.)

But religions are not persons, so unlike friends and spouses they do not count "in their own right" in an individualistic moral or political view. The only importance they have in such a view lies in their importance *for* those who accept them. My reasons for holding my religious views, however, may go beyond this subjective importance, for me, of being able to practice my faith, whatever it may be. If the reasons I see for holding my view claim to be not only conclusive reasons for *everyone* to hold this view, but also reasons why this view should be uniquely important in determining how everyone should live and what social policies and institutions we should have, then there is an incompatibility between this view and a tolerant political order. So we are back with the problem that Bruno Bauer described. This is parallel to the problem that arises for a person who believes his friend has unique moral importance, but such claims is more commonly made for religion than for friends.

The possible responses, however, seem to be the same as in the case of friends. One response is to claim that the problem will not arise as long as a religion (or comprehensive view) is reasonable in

the political claims that it makes. A second, more aggressive response is to claim, as we did in the case of friendship, that the only religious views that can form part of a person's good will be ones that are reasonable in this sense: ones that incorporate and are thus limited by, the demands of tolerance toward those who hold different views.

So Williams' problem and the problem faced by Marx and Rawls have the same structure. They are problems about how we can give a positive account of the ground of morality, or of justice, and the solution to them must lie not in the realm of morality or justice but in our accounts of individual good, and, more broadly, of what individuals have reason to want.

Contact Numbers

Conference Organizers

Kyla Ebels-Duggan: 847-477-4479
Mark Alznauer: 847-859-2217

Taxis

Evanston: There is almost always a queue of taxicabs at the corner of Orrington and Church in downtown Evanston, right outside the Orrington Hotel. But if you would like to call a cab:

Northshore Cab	847.864.7500
303 Taxi	847.556.0303
American Taxi	847.673.1000

Chicago: Cabs are easy to hail in most parts of Chicago. But here are a few listings:

American United Cab	773.248.7600
Yellow Cab	312.829.4222
Flash Cab	773.561.4444
Airport Express	773.247.1200

Department of Philosophy

Phone:	**847. 491.3656**
Fax:	**847. 491.2547**
Address:	**1880 Campus Drive** **Kresge Hall 2-345** **Evanston IL 60208**

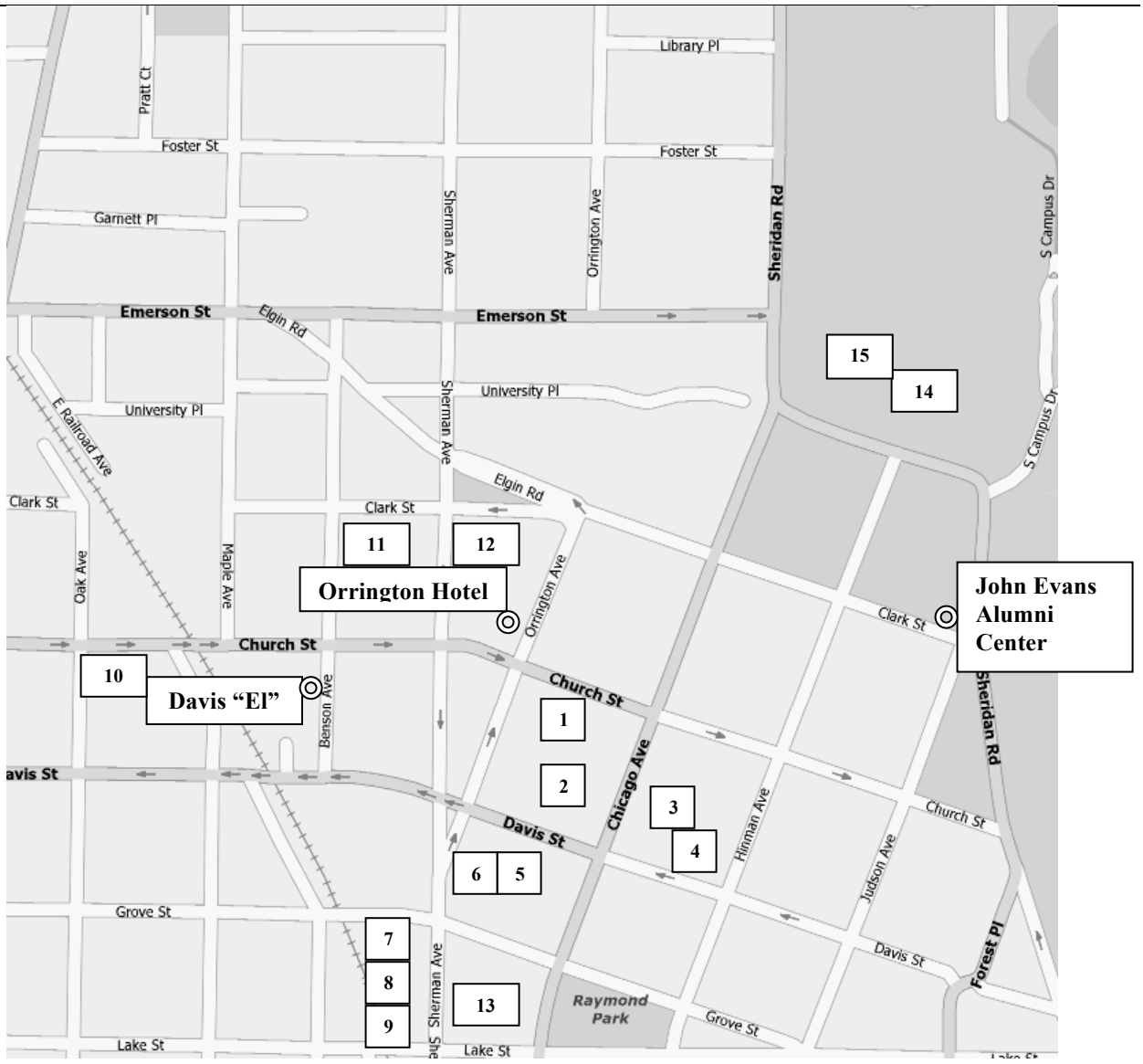
Reception Information

All are welcome to a reception on the ground floor of the Hilton Garden Inn.

1818 Maple Avenue, Evanston, Illinois, USA 60201
Tel: 1-847-475-6400 Fax: 1-847-475-6460

The reception will run from 6:30 to 10:30. Food and drinks will be provided by the hotel.

Map of Evanston



Places of Interest:

1 – Mt. Everest on Church St.
2 – Peet's Coffee on Chicago Ave.
3 – Tapas Barcelona on Chicago Ave.
4 – Davis Street Fish Market on Davis St.
5 – Koi Chinese and Sushi on Davis St.
6 – Potbelly Sandwiches on Orrington Ave.
7 – Bar Louie on Sherman Ave.
8 – Prairie Moon on Sherman Ave.
9 – Tommy Nevin's on Sherman Ave.
10 – Thai Sookdee on Church St.
11 – Merle's Barbecue on Benson Ave.
12 – Unicorn Coffee Shop on Sherman Ave.
13 – Best Western on Sherman Ave.
14 – Krege Hall, Philosophy Department
15—Harris Hall, History Department

Chicago Attractions

John Hancock Tower.

The best-kept secret in Chicago tourism is the Signature Lounge, located on the 96th floor of the Hancock Tower, 875 N. Michigan Ave. This bar/ restaurant provides guests with a 360 degree view of Chicago and Lake Michigan for the price of a drink -- there is no admission fee. Unlike observatory decks at the Sears and Hancock buildings, there are no lines.

The Magnificent Mile.

Chosen as one of the ten great avenues of the world, the Mag Mile is located just north of the loop and is Chicago's most prestigious shopping district. Water Tower Place, a very large mall, is located at 835 N. Michigan Avenue. Walking south on Michigan Ave. (or taking any of the many buses) you will end at the Wrigley Building down on the river (which you can follow into the loop and to Millenium Park and the Art Institute. It is especially pretty now with all the tulips.

Chicago Architecture Foundation Boat Tour.

$26 on weekdays (11 a.m., 1 p.m., and 3 p.m.), $28 on weekends (10 a.m., 11 a.m., 12 p.m., 1 p.m., 2 p.m., and 3 p.m.), 90 minutes long..
Dock location is southeast corner of the Michigan Avenue Bridge & Wacker Drive. Look for the blue awning marking the stairway entrance.
Buy tickets online: http://www.architecture.org/tour_view.aspx?TourID=8
Or call Ticketmaster: 312.902.1500 (reservations not required)

Millenium Park.

Millennium Park is located in the heart of downtown Chicago. It is bordered by Michigan Avenue to the west, Columbus Drive to the east, Randolph Street to the North and Monroe Street to the South. The park is open daily from 6 a.m. - 11 p.m. Admission is free. Attractions include the enormous mirror-surfaced bean sculpture, the Cloud Gate bridge, the Crown Fountains, the outdoor ampitheater and the Lurie Garden.

Chicago Art Institute.

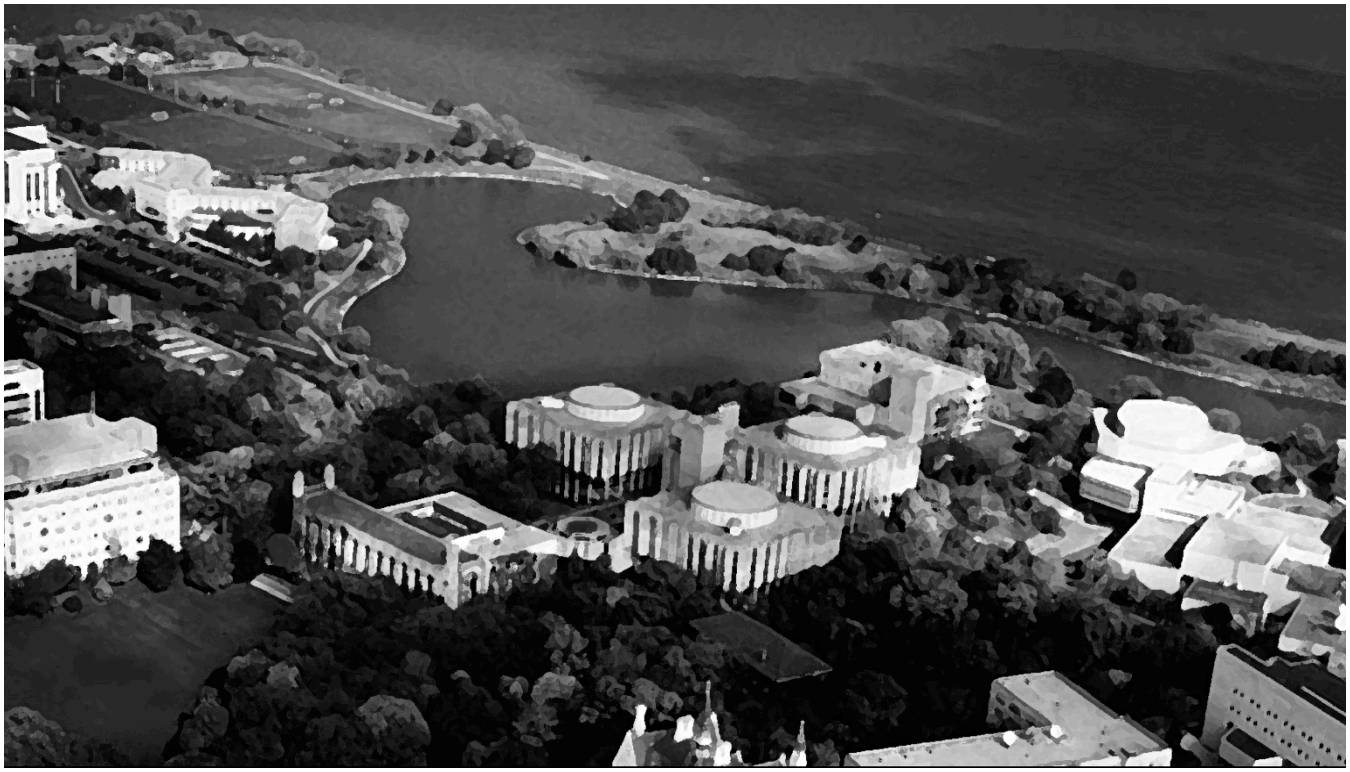
$12. Museum Hours: Monday --Wednesday, 10:30 a.m. -- 5:00 p.m. / Thursday, 10:30 a.m. -- 8:00 p.m. (Free 5:00 p.m. -- 8:00 p.m.) / Friday, 10:30 a.m. -- 5:00 p.m. / Saturday & Sunday, 10:00 a.m. -- 5:00 p.m.
Located at 111 South Michigan Avenue, Chicago, Illinois 60603-6404.
The closest L stop is at Adams/Wabash in the loop.
For more information: http://www.artic.edu/aic/visitor_info/

Shedd Aquarium.

Museum Hours: Weekdays 9 a.m. -- 5 p.m. / Weekends 9 a.m. -- 6 p.m.
Admission: Aquarium only -- $8 adults / All-access pass -- $23 adults and includes the Oceanarium, Wild Reef, Lizards and the Komodo King, Amazon Rising, the Caribbean Reef, Waters of the World.
To get to the Museum campus take the red line L to the Roosevelt stop, and board a museum trolley or take the #12 bus.

The Field Museum.

Museum Hours: Daily 9 a.m. -- 5 p.m. / Last admission at 4 p.m.
General Admission: $12 adults, up to $28 with extra exhibits
New Permanent Exhibit: The Ancient Americas.
Tickets available online: http://www.fieldmuseum.org
To get to the Museum campus take the red line L to the Roosevelt stop, and board a museum trolley or take the #12 bus.



Special thanks to those responsible for…

Content Providers:
Our conference speakers and commentators.

Conference Organization:
Mark Alznauer, Kyla Ebels-Duggan, Richard Kraut, Carlos Pereira Di Salvo, Raff Donelson, Lee Goldsmith, Seth Mayer, Wolfhart Totschnig, and Tyler Zimmer.

Faculty Paper Selection:
Mark Alznauer, Kyla Ebels-Duggan, and Richard Kraut.

Graduate Student Paper Selection:
Northwestern University Department of Philosophy Graduate Students.

Website Design:
Wolfhart Totschnig.

Administrative Support:
Crystal Foster, Emily O'Neill, and Tricia Liu.