Northwestern University Society for the Theory of Ethics and Politics

Eighth Annual Conference
March 13-15, 2014
Evanston, Illinois

Conference Program

TABLE OF CONTENTS

NORTHWESTERN UNIVERSITY
SOCIETY FOR THE THEORY OF ETHICS AND POLITICS

Eighth Annual Conference

March 13-15, 2014

Thursday, March 13, 2014

Morning Session:

9:00-10:25: "Must We Regret Weakness of Will?"

Mathieu Doucet, University of Waterloo

Commentator: Jesse Summers, Duke University

10:35-12:00: "Pragmatic Naturalism and the Functional Disunity of Ethics"

Nicholas Smyth, Brown University

Commentator: Daniel Trujillo, Northwestern University

Afternoon Session:

2:15-3:40: "The Specification Problem for Anderson's Democratic Egalitarianism"

Renee Bolinger, University of Southern California

Commentator: Jessica Talamantez, Northwestern University

3:50-5:15: "What is it to Accept a Promise?"

Abraham Roth, The Ohio State University

Commentator: Stephen White, Northwestern University

Friday, March 14, 2014

Morning Session:

9:00-10:25: "The Necessity of Duty and the Open Texture of Morality"

Jennifer Lockhart, Auburn University

Commentator: Heidi Giannini, Wake Forest University

10:35-12:00: "A Kantian Response to the Problem of Reception"

Mavis Biss, Loyola University, Maryland

Commentator: Daniel Groll, Carleton College

Afternoon Session:

2:15-3:40: "Cryptonormative Judgments"

Alex Worsnip, Yale University

Commentator: Raff Donelson, Northwestern University

3:50-5:45: Keynote Address - "What Are Theories of Desire Theories Of?"

Tamar Schapiro, Stanford University

Commentator: Anthony Laden, University of Illinois-Chicago

NORTHWESTERN UNIVERSITY
SOCIETY FOR THE THEORY OF ETHICS AND POLITICS

Eighth Annual Conference
March 13-15, 2014

Saturday, March 15, 2014

Morning Session:

10:35-12:00: "Risk and Motivation: Why 'What To Do?' Isn't Settled By 'What Should I Do?'"
Dylan Murray and Lara Buchak, University of California-Berkeley
Commentator: Debbie Goldgaber, Northwestern University

Afternoon Session:

2:15-3:40: "Wrongs, Rights, and Remedial Ambiguity"
Nicolas Cornell, University of Pennsylvania
Commentator: Gina Schouten, Illinois State University

3:50-5:45: Keynote Address - "On The Importance of Having a Life"
Gary Watson, University of Southern California
Commentator: Edward Hinchman, University of Wisconsin-Milwaukee

CONTACT INFORMATION

Conference Organizers:

Kyla Ebels-Duggan (Organizer): 847.477.4479
Richard Kraut (Organizer): 773.726.7742
Carlos Javier Pereira Di Salvo: 312.560.4554

Taxi Services:

Evanston:
Norshore Cab: 847.864.7500
303 Taxi: 847.556.0303
American Taxi: 847.673.1000

Chicago:
American United Cab: 773.248.7600
Yellow Cab: 312.829.4222
Flash Cab: 773.561.4444

Department of Philosophy

Phone: 847.491.3656
Fax: 847.491.2547
Address: 1880 Campus Drive
Kresge Hall 2-345
Evanston, IL 60208

NORTHWESTERN UNIVERSITY
SOCIETY FOR THE THEORY OF ETHICS AND POLITICS

Eighth Annual Conference
March 13-15, 2014

RECEPTION INFORMATION

All speakers and commentators are welcome to attend a buffet dinner on Friday evening following the afternoon session.

Location: John Evans Alumni Center
1800 Sheridan Road
Evanston, IL 60208

The reception will last from approximately 6:00 to 10:00 in the evening.

NORTHWESTERN UNIVERSITY
SOCIETY FOR THE THEORY OF ETHICS AND POLITICS

Eighth Annual Conference

March 13-15, 2014

A SPECIAL THANKS...

CONTENT PROVIDERS:
OUR CONFERENCE SPEAKERS AND COMMENTATORS.

CONFERENCE ORGANIZERS:
KYLA EBELS-DUGGAN, RICHARD KRAUT, STEPHEN WHITE, CARLOS PEREIRA DI SALVO, DAN SKIBRA

FACULTY PAPER SELECTION:
KYLA EBELS-DUGGAN, RICHARD KRAUT, STEPHEN WHITE

WEBSITE DESIGN AND MAINTENANCE:
DAN SKIBRA

ADMINISTRATIVE SUPPORT:
CRYSTAL FOSTER, JASMINE BOMER, TRICIA LIU, ERIC BARRONE

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.

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).

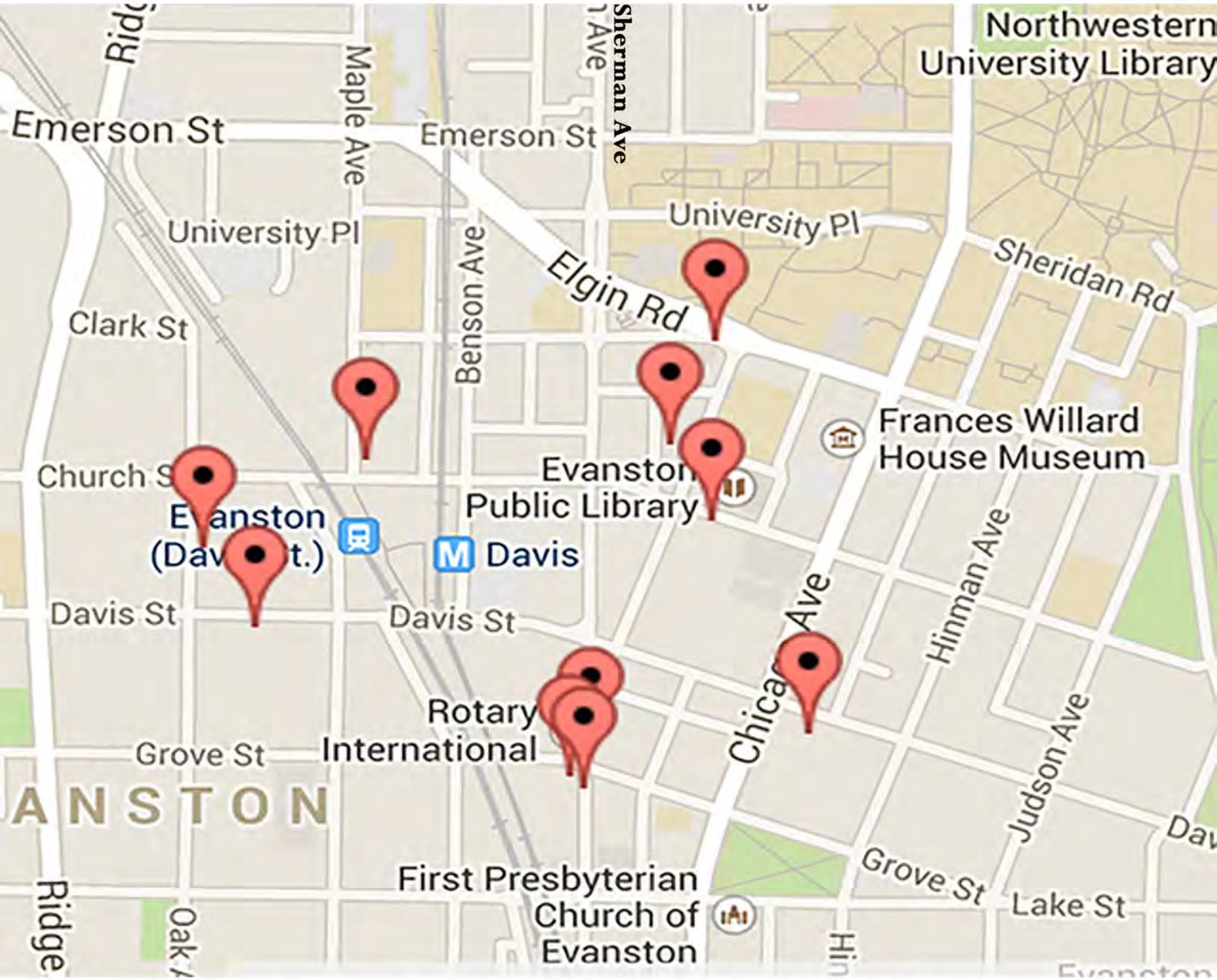
Chicago Architecture Foundation Boat Tour:
$26 on weekdays (11am, 1pm, 3pm) and $28 on weekends (10am, 11am, 12pm, 1pm, 2pm, and 3pm), 90 minutes long. Dock location is southeast corner of the Michigan Avenue Bridge and Wacker Drive. Look for the blue awning marking the stairway entrance. You can buy tickets online.

Millennium 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. This park is open daily from 6am to 11pm. 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.

Shedd Aquarium:
Museum Hours: Weekdays: 9am-5pm & Weekends: 9am-6pm.
Admission: $8 adults for aquarium only, $23 for all-access pass that includes Oceanarium, Wild Reef, Lizards and the Komodo King, Amazon Rising, the Caribbean Reef, Waters of the World, and others. To get to the museum, take the red line L to the Roosevelt stop and board a museum trolley or take the #12 bus.

The Field Museum:
Museum Hours: 9am-5pm. General admission is $12 for adults and up to $28 with extra exhibits. Take the red line L to the Roosevelt stop and board a museum trolley or take the #12 bus.

MAP OF EVANSTON



27 Live [1014 Church, 855-927-5483]
Eat, Drink, and Listen.

Bar Louie [1520 Sherman, 847-733-8300]
Neighborhood Bar and Grill.

Celtic Knot [626 Church, 847-864-1679]
Irish Public House and Restaurant.

Cosi [1740 Sherman, 847-328-2050]
Sandwiches, Soups, and Salads.

Giordano's Pizza [1527 Chicago, 847-475-5000]
Best Pizza in Town.

Joy Yee Noodle [521 Davis, 847-733-1900]
Pan Asian.

La Macchina [1620 Orrington, 847-425-1080]
Good Food, Shared Plates, Wine Bar, Locally Roasted Coffee.

Mt. Everest [630 Church, 847-491-1069]
Royal Indian and Nepakese Cuisine.

Hilton Orriginton Hotel [1710 Orrington, 847-866-8700]
One block west of Chicago Avenue, near campus.

Best Western Hotel [1501 Sherman, 847-491-6400]
South on Sherman Ave.

Mathieu Doucet
University of Waterloo

Mathieu Doucet is Assistant Professor in the Department of Philosophy at the University of Waterloo. His research is in ethics, with an emphasis on moral psychology. His current research concentrates on weakness of will, hypocrisy, and the moral significance of self-knowledge.

"Must We Regret Weakness of Will?"
Commentator: Jesse Summers, Duke University

Abstract: *The two dominant models of weakness of will disagree about a great deal, including whether it involves the violation of a judgment or of an intention, whether or not weak-willed agents act contrary to an intention or a judgment they hold at the time of action, and whether it has a normative component.*

In this paper, I argue that a) despite these disagreements, both models are united by the assumption (either implicit or explicit) that regret is a typical or even necessary element of standard cases of weakness of will, and that b) this assumption is mistaken. I draw on empirical and philosophical work on self-assessment to show that regret need not accompany typical weak-willed behaviour. I then conclude by arguing that abandoning the regret condition forces us to revise our understanding of the nature of weakness of will.

1. Introduction

Those of us who suffer from weakness of will often regret doing so. Consider an example: I resolve not to drink at the department Christmas party. Once at the party, though, my colleague breaks out a new bottle of Laphroig, and I find myself indulging despite my earlier resolution. The next morning, head pounding, I reproach myself for weakly giving in to temptation and breaking my resolution.

This is an example that any philosophical model of weakness of will should treat as a paradigm case, and one reason for this that my regret seems to be evidence that abandoning my resolution was a weak-willed. My action seems weak not just because I do what I earlier resolved not to do, but because I myself realize that doing so was a mistake, and reproach myself for my own failure.

In this paper, I argue for three main conclusions. First, I show that several otherwise divergent accounts of weakness give regret a central role. Second, I draw on psychological research on self-assessment to argue that this is a mistake, and that regret actually plays a much less of a role in weakness of will than is typically assumed. Finally, I argue that combining these two claims forces us to reconsider the dominant account of weakness of will, and to give a different explanation of what makes weakness of will (typically) irrational.

2. Regret

Regret is an emotion that arises in many contexts, not all of them relating to weakness of will. The form of regret connected to weakness of will has several characteristics. First, if I regret my drinking the whisky, then at minimum, I judge I did something I ought not to have done. Second,

I judge that it would have been better for me *not* to have done it.[1] Third, my regret involves a particular self-assessment: I judge that the *cause* of my doing something I ought not to have done is weakness of will.[2] I can regret doing something wrong as a result of ignorance, or malice, and this will involve a different judgment than that involved in regretting weakness of will. Finally, this judgment will constitute (or be accompanied by) a negative emotional reaction. Regret is not just a judgment about the causes of my action, it is also a form of self-reproach: a self-directed reactive attitude.

This means that, *if* it is true that weak-willed agents typically experience regret, then this will be because (1) weak-willed agents typically engage in retrospective self-assessment, *and* (2) when they do, they typically form *accurate* judgments. Below, I argue that both of these are false. First, however, I show that several leading models of weakness of will are committed to the idea that they are true.

3. Models of weakness of will

3.1 Akrasia: It is helpful to divide philosophical models of weakness of will into to broad categories. The long dominant model describes it as intentional action contrary to the agent's considered or 'best' judgment.[3] Common to the various versions of this *akratic* model is that those who act akratically act contrary to the judgment they hold *at the time* about what they have most reason to do. Weakness of will is puzzling, on this view, because it makes us act irrationally *by our own lights* at the very time of action.

This view of the nature of weakness of will means that weak-willed agents tend to have a kind of self-knowledge: because they act contrary to what they still consciously judge to be best, they *know* they are weak-willed.[4] This makes regret seems almost inevitable, and is what leads Aristotle to distinguish akratic from intemperate agents on the grounds that only akratics are "prone to regret".[5]

3.2 Judgment-shift: Recently, some philosophers have argued that weak-willed agents do *not* typically act contrary to their own considered judgments—that is, they don't tend to judge, at the time of action, that they are failing to do what they have most reason to do. Instead, a much more common form of weakness of will involves an over-ready *revision* of our judgments about what we have most reason to do. Such philosophers offer otherwise very different models of weakness

[1] This second condition is necessary because it will be false in cases of tragic dilemmas, where the kind of regret present is quite different.

[2] . This need not mean that I employ the concept 'weakness of will', or have a particular philosophical model in mind. It simply means that I believe that I succumbed to temptation or violated my own prior commitment.

[3] The first version of this model is Aristotle's, particularly in *Nicomachean Ethics* Bk. VII. It was revived by Donald Davidson's "How is weakness of the will possible?" in Joel Feinberg, *Moral Concepts* (Oxford University Press, 1970). Perhaps it most prominent contemporary defender is Alfred Mele, in, for example, *Irrationality* (Oxford University Press, 1987) and "Weakness of will and akrasia", *Philosophical Studies* 150 (2010), pp. 394-404.

[4] Dylan Dodd's account of *akrasia* does not require the violation of a consciously entertained commitment, so long as the action violates a policy the agent "continues to have at the time (s)he performs the action." For Dodd, however, whether a policy remains in place is revealed by future regret. " Weakness of Will as Intention Violation", *European Journal of Philosophy* 17 (2007) pp. 45-59 at 48.

[5] Aristotle, *Nicomachean Ethics* 1150b30.

of will, but they are unified in seeing it, not as action contrary to one's best judgment, but as action brought about by a weak or irrational *shift* in judgment.[6]

The clearest cases of this form of weakness of will involve what Richard Holton, following Michael Bratman, calls "resolutions": intentions "designed to stand firm in the face of contrary inclinations."[7] To return to the whisky: perhaps the morning of party I decided not to drink. In the morning I generally have no strong desire for Islay whisky, but I have enough self-knowledge to realize that in the evening, at parties, I often *do* have such a desire. I therefore *resolved* not to drink: I committed to not reconsidering or revising my intention, even in the face of future temptation. If, at the party, I reconsider and succumb to temptation, then I have reopened an issue I had already decided was closed, *and* treated my desire to drink Laphroig as a reason to drink it, even though I had previously resolved to ignore such desires. So on this model if, under the sway of Laphroig's peaty aroma I change my mind and take a drink, I am weak-willed. This is so even if I deliberate prior to drinking and offer a rationalizing justification—I might say "I didn't know there'd be a bottle of the 18-year-old! And just one dram won't hurt. I was too pessimistic this morning". Offering such a justification at the time of action need not, on the judgment-shift view, rescue me from weakness of will.

An important difference between the two models is therefore that, unlike the akratic model, the judgment shift model involves no internal inconsistency, at the time of action, between what I do and what I judge I have most reason to do. Instead, my will is weak because I have too-readily *changed* my judgment. Moreover, since there is no internal inconsistency, I cannot be *aware* of it, and so do not have any self-knowledge of my own weakness.

Because this model does not identify weakness of will with internal inconsistency, it faces a challenge that the akratic model does not: how can we distinguish between weak-willed intention-revision and perfectly rational changes of mind? Intentions are supposed to be stable and put and end to deliberation,[8] but they are also supposed to be defeasible—changing one's mind is not always a sign of irrationality. Sometimes, we really *ought* to reconsider and revise our intentions, even if they were rational at the time that we formed them. While different judgment shit models explain this distinction is different ways, *regret* plays an important role in many of them.

4. Judgment-shift and regret

I suggested above that regret seems to be almost a defining feature of the akratic model. But things might appear different with the judgment-shift model, since unlike the akratic model it does not require, and in fact rules out, the recognition of an internal inconsistency, at the time of action, between judgment and action. And if weak-willed agents don't *know* they've acted weakly, they cannot regret doing so.

---

[6] Examples include Richard Holton, "Intention and weakness of will", *Journal of Philosophy* 96 (1999) pp. 241-262; *Willing, Wanting, Waiting* (Oxford University Press, 2009); Alison McIntyre, "What is wrong with weakness of will?", *Journal of Philosophy* 103 (2006), pp. 384-311; Michael Bratman, "Planning and Temptation" in *Faces of Intention* (Cambridge University Press, 1999), pp. 35-57, and "Temptation revisited" in *Structures of Agency* (Oxford University Press, 2007), pp. 257-282; George Ainslie, *Breakdown of Will* (Oxford University Press, 2001); and Neil Levy, "Resisting weakness of will", *Philosophy and Phenomenological Research* LXXXII (2011) pp. 134-155.

[7] Holton, *Willing, Waiting, Wanting* p. 10.

[8] A point both Holton and Bratman repeatedly emphasize.

Remember, however, that the judgment-shift model owes us an explanation of the difference between weakness of will and (mere) changes of mind. To give an account of why abandoning one's resolution is weak-willed, it seems that we need to have an explanation for why it would be better to maintain one's resolution, and so rule out cases in which reconsideration would be rational. This can be difficult: why should what I wanted in the *past* overrule what I want *now?*

4.1 Normative Models: One way to answer this question is *normative*—it is to offer an account of when it would be *rational* to change one's mind, and then identify weakness of will with irrational shifts in judgment. Regret plays a central role in such accounts: for example Holton's account of the rationality of intentions is largely based on the work of Michael Bratman.

Bratman's account of the conditions under which maintaining one's prior intentions is rational: "(a) if you stick with your prior intention, you'll be glad you did it; and (b) if you do not stick with your prior intention, you will wish you had."[9] It is rational to stick to one's resolutions provided that both conditions are met, and (b) is a regret condition—a resolution is rational if you would regret breaking it. This account of the rationality of resolutions indirectly gives us an account of weakness of will where regret is central. If persisting with a resolution is rational just in case failure to stick to it would lead to regret, then *abandoning* resolutions is irrational just in case doing so would *lead to* regret. It is therefore the potential presence or absence of regret that distinguishes weakness of will from mere changes of mind.

If we accept this account of the rationality of stable intentions and resolutions, then it seems we are driven to the view that what makes a shift of judgment weak-willed is that it is regretted after-the-fact: in other words, that the agent engage in retrospective self-assessment.

We can see a similar argument in the work of Harry Frankfurt, who considers the possibility that "someone who has decided to perform a certain action may discover, when the chips are down, that he just cannot go through with it."[10] Frankfurt has in mind something like an akratic inability to act on one's intention, but his aim is to distinguish two different sources of such an inability. On the one hand, an agent may be unable as a result of psychic forces that are "not in the fullest sense his own… and whose influence he struggles to resist."[11] This sort of aversion sounds quite close to irrational *akrasia*. On the other hand, the agent may *endorse* the aversion that prevents carrying out his intention, and see it as "in the most authentic sense his own force."[12] Frankfurt's point in drawing this distinction is that not all cases in which someone finds himself unable to act on his judgment are irrational, since "a person's feelings may accord better with reason than his judgment does."[13] What sets irrational weakness apart from rational wholeheartedness is the agent's overall assessment of the aversion—his endorsement or rejection of the psychological force blocking him from acting on his judgment. While Frankfurt's main aim here is not to explain the nature of weakness of will, but rather to explain the substantive

[9] Bratman 1999, p. 79.
[10] Harry Frankfurt, "Rationality and the unthinkable" in *The Importance of What We Care About* (Cambridge University Press, 1998) p. 182.
[11] Frankfurt, p. 183.
[12] Frankfurt . p 184.
[13] Frankfurt., p. 189. For other versions of the claim that something like *akrasia* can be rational, see Nomy Arpaly, "On acting rationally against one's best judgment", *Ethics* 110 (2000), pp. 485-513; and Robert Audi, "Weakness of will and rational action", *Australasian Journal of Philosophy* 68 (1990), 270-281.

content of the will of rational agents, his view shares a commitment to the idea that the agent's own assessment of the failure to act on his or her intention is central in determining whether such a failure counts as irrational or weak-willed.

4.2 Descriptive models: The approaches described above are based on a view of what changes of mind would be *rational*, and so have a normative dimension. But there are also descriptive models that identify weakness of will with judgment-shifts brought on by particular psychological mechanisms.

Neil Levy has recently argued that the shifts in judgment we call weak-willed involves changes of mind caused by a shift from System 2 to System 1 cognitive processing brought on by ego-depletion.[14] Levy is making deflationary argument—since the same mechanism can cause shifts we are *not* inclined to describe as weak-willed, he argues that weakness of will is not a genuine psychological kind. Nevertheless, his argument that the proposed mechanism aptly describes seemingly weak-willed judgment shifts relies essentially on the presence or absence of regret.[15]

Finally, George Ainslie explains weakness of will as the result of changes of mind brought on by temporary preference inversions caused by the hyperbolic discounting of the value of future goods.[16] As he puts it, the puzzle with weakness of will is explaining how agents do something "while knowing they'll regret it and even while trying to stop".[17] So built into the very fabric of weakness of will is the idea that the judgment shift is *temporary*, and that it shifts *back*. Indeed, on Ainslie's view that such reversals are "avoided if foreseen… and regretted afterwards" is a "defining feature" of weakness of will.[18] While Levy and Ainslie's descriptive models explain weakness of will by appeal to very different psychological mechanisms, both of them identify the target of those explanations with essential reference to a form of regret.

In sum, the evidence surveyed suggests that a wide variety of models of weakness of will—both akratic and judgments shift, and both normative and descriptive—build into their accounts the idea that weak-willed agents typically look back on their decisions with regret, and accurately see that they acted weakly. But we have good reason to believe that this rarely happens.

5. Against regret

If the regret associated with weakness of will is characterized by an accurate retrospective self-assessment of the causes of the agent's behaviour, then the question is whether weak-willed agents do, in fact, typically engage is such clear-eyed self-assessment. There are good reasons for doubting it, both because there are good reason reasons for doubting that *any* of us typically do so, *and* because there are additional reasons for supposing that agents susceptible to weakness of will are particularly unlikely to be accurate self-assessors.

---

[14] Levy, "Resisting weakness of will".

[15] As he puts it, "Genuinely weak judgment shifts are relatively brief and transitory; once the depleted resources are restored, the agent typically regrets the action… It is precisely because changes of mind are produced through deliberation and are not regretted by agents that we— rightly— do not regard them as involving failures of practical rationality" Levy, "Resisting weakness of will" p. 139.

[16] George Ainslie, *Breakdown of Will* Part 1.

[17] Ainslie, p. 51.

[18] Ainslie, p. 49.

There are two broad ways in which an agent can be weak-willed and yet fail to experience regret. First, an agent might engage in active, but mistaken, self-assessment, and so fail to conclude that her action was weak-willed. Second, she might simply decline to engage in such self-assessment to begin with.

5.1 Mistaken self-assessment: The first possibility is given support by evidence from psychology. If the regret characteristic of weakness of will depends on accurate self-assessment, then it requires agents to accurately recall the beliefs, desires, intentions, and reasoning that led to their actions. Such recall is in fact difficult to achieve: there is ample evidence that we are quite bad at recalling both our former intentional attitudes, *and* the cognitive processes led us to hold our current attitudes. Much of this evidence comes from studies on the ways we explain our own behaviour, which has clear relevance in the context of weakness of will.[19] When asked to explain why we acted as we did, it *seems* to us (in standard cases) that we answer by engaging in introspection, and *remember* the reasoning processes that led to our actions. In fact, the evidence suggests that such cognitive processes are quite often unavailable to introspection. Instead, what we often do is rely on what we *take* to be a plausible general theory of behaviour, which we mistakenly take to be recall of the reasoning the led to the behaviour. So when I explain why I did what I did, it *seems to me* that I'm remembering, but I'm often confabulating a plausible *post hoc* story.

The problem here is that this process is less than fully reliable. This is because what we *take* to be a plausible theory is quite often false, either in general or in our own case. And we can see a straightforward way in which this *post hoc* problem can prevent a weak-willed agent from feeling regret. If an agent's action is weak-willed, but he expects that such behaviour is *not* typically the product of weakness of will, then in attempting to recall his motives he may form a false belief about what motivated his actions, and so mistakenly deny that his action was weak-willed.

The problem for the judgment shift model is even deeper than this, however. It's not just that we can fall into error about why we acted. More relevant to the argument here is the discovery that people who change their judgment will often sincerely deny having done so, and so will, after changing their minds, believe that they always held their new judgment. This will clearly block weak-wiled agents from accurately recognizing that their change of judgment was weak-willed, since they *will not know* that there was any change to assess.

In many recall studies, subjects are questioned on their attitudes, then convinced (in various ways) to *change* their attitudes and finally, asked to engage in a form of active self-assessment by *recalling* previous attitudes and comparing them to their current attitudes.[20] Even when prompted to engage in such self-assessment—in fact, even when explicitly reminded that

[19] The classic source is Richard Nisbett and Timothy Wilson, "Telling more than we can know: verbal reports on mental processes", *Psychological Review* 84 (1977) p. 231. A more recent defense of the view, applied to moral reasoning in particular, is Jonathan Haidt's "The emotional dog and its rational tail: a social intuitionist approach to moral judgment", *Psychological Review* 108 (2001) p. 814-834.

[20] The classic references are Daryl Bem and Keith McConnell, "Testing the self-perception explanation of dissonance phenomena: on the salience of premanipulation attitudes" *Journal of Personality and Social Psychology* 14 (1970) 23-31; and George Goethals and Richard Reckman , "The perception of consistency in attitudes" *Journal of Experimental Social Psychology* 9 (1973) pp. 491-501 (both cited in Nisbett and Wilson 1977). The effect has also been found for the recall of emotions, as in Linda Levine and Martin Safer, "Sources of bias in memory for emotions" *Current Directions in Psychological Science* 11 (2002), pp. 169-173.

the experimenter had a record of the subjects' original attitudes— subjects perform poorly. Recall of past attitudes is biased toward present experience: people who change their minds often believe that they have *always* believed what they *currently* believe. This suggests that when we change our minds, we are often quite unaware of having done so, even when we engage in active introspective self-assessment.

This is a serious challenge to the idea that weakness of will leads to regret: if regret requires, at a minimum, *knowing that you changed your mind*, then in many cases, the weak-willed lack this knowledge. So the well-established present-experience bias in recall can serve to prevent regret from gaining a purchase, since even if we do engage in self-assessment it can prevent us from recognizing that we changed our minds, abandoned our resolutions, and acted weakly.

The tendency to 'recall' previous motivations in a way that lines up with expectations connects to another well-established set of psychological biases: both the tendency of agents to interpret their behaviour in a way that is consistent with their self-image, *and* bias most people display in favour of a *positive* self-image.[21] Both of these biases can serve to militate against accurate self-assessment on the part of the weak-willed agent: even if we *are* aware of having changed our minds, these biases can lead us to interpret the rationality of our shift in a way that a) preserves our sense of internal consistency, and b) preserves our positive sense of self. Again, here we can see how such biases can lead the weak-willed agent to evaluate the rationality of his or her judgment-shift and conclude that nothing untoward occurred.

Finally, agents who retrospectively assess the rationality of their decisions tend to engage in "post-decision consolidation through re-evaluation".[22] That is, we tend to confirm the rationality of our decisions by reevaluating the options and lowering our assessment of the non-chosen option, making the decision seem even more straightforward than it was at the time. When we do so, our retrospective assessments are unlikely to be that our decision was irrational or weak-willed, and so this process will also militate against regret in cases of weakness of will.

5.2 Avoiding self-assessment: The argument thus far has pointed out that, *even if we do engage* in active self-assessment, it is not guaranteed to be accurate. But there is a second way that the weak-willed agent can fail to experience regret: she might simply fail to engage in retrospective self-assessment at all. In fact, several of the biases just mentioned can get a grip on our beliefs by blocking reconsideration of the rationality of our decision—if we are biased in favour of a coherent or a positive self-image and have yet to re-assess our actions, then we can preserve a positive assessment by simply not engaging in *re*-assessment.

In fact, on Holton's view a general tendency not to reconsider is an essential component of *strength* of will: for someone to be rational with respect to resolutions is for them to exhibit a

[21] For biases in favour of positive self-assessment, see, for example, Mark D. Alicke, "Global self-evaluation determined by the desirability and controllability of trait adjectives" *Journal of Personality and Social Psychology* 49 (1985) pp. 1621-1630; Ziva Kunda, "The case for motivated reasoning", *Psychological Bulletin* 108 (1990) pp. 480-498; and Justin Kruger and David Dunning, "Unskilled and unaware: how difficulties in recognizing one's incompetence lead to inflated self-assessments" *Journal of Personality and Social Psychology* 77 (1999) pp. 1121-1134.

[22] Ola Svenson and Lars Benthorm, "Consolidation processes in decision making: post-decision changes in the attractiveness of alternatives", *Journal of Economic Psychology* 13 (1992), pp. 315-327.

tendency not to reconsider that it would be rational to have. This builds in a normative dimension, since a tendency not to reconsider can be irrational if it is too strong: someone who *never* reconsiders exhibits an irrational tendency (and pig-headed stubbornness). So it's only if the tendency not to reconsider is properly calibrated that it can generate rational strength of will. My point here is that a tendency not to reconsider can also contribute to irrational *weakness* of will if it is combined with the sorts of biases or motivational irrationalities that lead to weak-willed judgment shifts.

Indeed, it is possible that the kinds of motivational biases and irrationalities that generate failures of self-assessment are often closely related to the very biases and irrationalities that generate weakness of will. The judgment shifts characteristic of weakness of will are irrational in part because they reveal a susceptibility to motivated rationalization, and it is the exact same susceptibility—often with very similar motivations—that generates the mistaken self-assessments that follow weakness of will. And if both weak-willed judgment shifts and mistaken self-assessments have a common cause or are generated by a similar psychological mechanism, then this gives us good reason to suspect that regret will *not* typically result from weakness of will.

6. Conclusion

Both akratic and judgment shift models of weakness of will draw a close connection with regret. In fact, judgment-shift models have often explained weak-willed judgment shifts by appeal to regret. If the argument offered in the last section is correct, however, this is a mistake, since we have no reason to suppose that there is such a close connection. The conclusion we ought to draw, however, is *not* that the judgment-shift model is mistaken. In fact, since regret involves a *retrospective* assessment, many of the arguments offered in the previous section apply to the akratic model as well. Rather, the conclusion to draw is that judgment-shift models need to give more thought to the normative dimension of the rationality of intention-revision.

For example: descriptive models cannot simply argue that they've hit upon the right mechanism because it is one that generates regret or is correlated with regret—if the argument above is right, then the mechanisms that generate weakness of will might very well *also* militate against regret. Any descriptive model, then, should justify its proposed mechanism not on the grounds that it explains regret, but rather that it accurately captures irrationally weak shifts in judgment. And this suggests that even accounts that rely on descriptions of psychological mechanisms should have a normative component.

More purely normative accounts should also be modified. It is not enough to say that what makes a shift irrational is that it leads to regret: this leaves the assessment of the rationality of the shift at the whims of particular individual judgments that themselves might be the product of motivated biases. Nor does it help to say that shifts of judgment that would be regretted *in the absence of bias* are weak-willed, at least if both the initial judgment-shift and the failed self-assessment can have a common cause. After all, we're dealing with agents who are clearly irrational to some extent—expecting that their irrationality is narrowly confined is unrealistic. This suggests the problem with relying on the presence or absence of regret to determine whether a change of mind is weak-willed: it involves assessing the rationality of an agent's change of mind by deferring to the assessment of the agent in question, in cases where the agent has just revealed himself to be suffering from a significant level of practical irrationality. It seems unlikely that such assessments will be reliable.

A convincing account of weakness of will must therefore navigate a tricky middle path. It should preserve the sense that weakness of will involves a failure *by the agent's own lights*, but it should not understand this in terms of the agent's own particular in-the-moment assessment. This suggests that a plausible approach will involve a holistic evaluation of the agent's overall (internal) reasons for acting.[23] If this is right, then an agent's shift in judgment could genuinely be weak-willed by the agent's own standards, without the agent himself judging it to be so in hindsight. Such a model would walk the required middle path, and so would break the close connection between weakness of will and regret. But developing the details this model is a task for another time.

[23] See for example, Audi, "Weakness of will and rational action".

Nicholas Smyth

Brown University

Nicholas Smyth is a graduate student at Brown University. He works on value theory, broadly construed. In meta-ethics, he studies the intersection of expressivist and pragmatist theory. His work on normative ethics focuses on Kant, Nietzsche, and Bernard Williams, and he occasionally makes forays into ancient philosophy.

"Pragmatic Naturalism and the Functional Disunity of Ethics"
Commentator: Daniel Trujillo, Northwestern University

Abstract: *It has recently become common for naturalistically-minded moral philosophers to provide genealogies of human morality in order to defend views about its function [Gibbard (1990), Joyce (2006). Philip Kitcher's The Ethical Project is perhaps the most ambitious of these attempts, since it purports to provide an entirely new meta-ethical framework under which this form of genealogical inquiry can proceed. Kitcher labels this framework pragmatic naturalism, and he claims that it delivers the conclusion that the basic function of ethics is to "remedy altruism failure". Furthermore, he claims that our own moral progress depends crucially on recognizing that this 'ur-problem' can receive new and better solutions given changing cultural circumstances. In this paper, I show that the real history of human ethical practice generates several empirical and philosophical hurdles for Kitcher. I argue that this history, combined with his own view of what a function is, generates three alternative functions that ethics might plausibly play in human life. I further argue that each of these functions grounds a powerful normative challenge that threatens to undermine Kitcher's project. I conclude that human ethical practice is very likely characterized by what I will call functional pluralism. The more general lesson is that the pragmatic naturalist must not seek shelter from the messy contingencies of human history. A clear-eyed investigation into the history of morality will probably not uncover a single function which can enable us to articulate a simple notion of moral progress.*

> The cause of the origin of a thing and its eventual utility, its actual employment and place in a system of purposes, lie worlds apart.
>
> – Nietzsche, *Beyond Good and Evil*

A perennial criticism of analytic philosophy is that its methods are ahistorical.[24] However, it has recently become common for naturalistically-minded philosophers to provide genealogies of

[24]Richard Rorty, for example, accused analytic philosophers of "trying to escape from history." See Rorty, R., (1980). *Philosophy and the Mirror of Nature*. Princeton, NJ: Princeton University Press, p. 8.

morality in order to defend views about its function.[25] Perhaps the most ambitious of these philosophers is Phillip Kitcher. Kitcher has outlined and defended a methodological approach he labels *pragmatic naturalism*, a philosophical program which eschews rationalistic attempts to provide a priori grounding for ethical precepts, and which turns to actual human history to unearth the constitutive function of moral practice. To this end, Kitcher has drawn extensively on the anthropological record to provide support for his claim that the original function of morality is to promote social cohesion by "remedying altruism failure."[26] Functional accounts like Kitcher's promise to shed new light on the content and justification of moral norms, and they represent a move away from the days of ahistorical conceptual analysis. While I have considerable sympathy for pragmatic naturalism, my view is that Kitcher is far too quick in drawing his conclusions, and *increased* attention to history is required if we are to appreciate this fact. In this paper, I take Kitcher as my primary target, but I suspect that many of the objections I will raise can be pressed against other historically-oriented moral philosophers who believe they have identified the function of morality.[27]

In what follows, I will attempt to mirror Kitcher's own reasoning about history and function in order to identify three basic problems with his account. The first is that he ignores the possibility that morality has *individual*-level functions which predate and even conflict with the group-level function he identifies. The second is that the processes of group-selection which are alleged to have formed the ethical project will likely yield a much more disunified picture of morality's original function(s). The third is that Kitcher does not try to ascertain whether conditions that have obtained since the end of the Pleistocene era may have altered, subverted or even undermined the original function(s) of morality. Each of these empirically plausible scenarios creates a distinctive kind of *normative* challenge for Kitcher, one grounded in his own meta-ethical commitments. Before proceeding to these criticisms, however, it will be necessary to outline the conception of functions that fuels Kitcher's project.

---

[25] For example, the expressivists Alan Gibbard and Simon Blackburn believe that the evolution of morality shows that its function is, roughly, to produce co-ordination in human populations. Gibbard, Allan, (1992) *Wise choices, apt feelings*: A Theory of Normative Judgment. Harvard University Press. Blackburn, Simon, (1998). *Ruling Passions*. Oxford: Clarendon Press. Joyce, Richard (2006). *The Evolution of Morality*. The MIT Press. Elsewhere, Robert Nozick claimed, on evolutionary grounds, that the central function of ethics is "cooperation to mutual benefit." More recently, Neil Sinclair argues that morality functions so as to "[allow] groups of interacting individuals to co-ordinate their actions and emotions for mutual benefit." And Eric Campbell, in a more pessimistic vein, argues that morality has the function of producing "commitment strategies" which produce a "deflection of attention from our motivations and values." See Nozick, R. (2001). *Invariances: The Structure of the Objective World*. Harvard University Press, p. 247-267, and Sinclair, N. (2012), "Metaethics, Teleosemantics and the Function of Moral Judgements," *Biology & Philosophy*, *27*(5), 639-662, and Campbell, Eric (2014), "Breakdown of Moral Judgment," forthcoming in *Ethics*, April 2014.

[26] Kitcher, Philip, (2011). *The Ethical Project*. Harvard University Press.

[27] I lack the space to develop this thought in this paper, but if the reader is interested, he or she may want to bear in mind that metaethical expressivists often draw on evolutionary history in order to defend claims about the primary function of moral discourse. Unlike Kitcher, these philosophers do not seek to draw any substantial, first-order conclusions from premises about the function of morality, but they nonetheless believe that moral discourse is fundamentally *for* coordination, in the biological sense. This claim is particularly vulnerable to the first objection that I will raise against Kitcher. Biological functions can appear at many levels, and expressivists need to defend their exclusive focus on group-level functions if they are to make good on their claims.

1. Kitcher's Theory of Function

Function, for Kitcher, is fundamentally a result of *design*. "The function of S," he writes, "is what S is designed to do".[28] Thus, the idea of a selection-pressure plays a crucial role in his theory. Selection pressures, taken together, form what he calls a "problem background", and design occurs when a solution to that problem background appears. These solutions rest on a continuum, from the "blind" processes of natural and cultural selection (the oft-cited "design without a designer") to conscious problem-solving by intentional agents.

One complication for any view that derives functions from selection-histories arises from the fact that "selection-pressures" can and do occur at multiple *levels*. Thus, honeybees may have wings because wings enabled particular individual bees to survive and reproduce, but they may have stingers because a group of bees with stingers is much more capable of defending its hive. Even though acts of stinging are often fatal to individual honeybees, the trait evolved because it promoted *group* survival. In analyzing the evolutionary functions of morality, it is important that we bear in mind that it may have both individual-level functions and group-level functions.

It will also be important to mention another feature of these sorts of accounts. When they are used to analyze human practices or dispositions, it is highly likely that they will uncover functions of which we are not normally conscious when we engage in the relevant practice or display the relevant disposition. We normally associate the word "function" with conscious intention or purpose, but Kitcher's account posits no necessary connection, here. The function of a thing is given by what is has been designed to do, and not by what we *mean* for it to do, though of course these can coincide (as in the case of consciously designed artifacts). Philosophers who perform historical analyses of the function of morality should be prepared to discover that our moral practices serve functions of which we are not normally aware.

Bearing these methodological qualifications in mind, we can see that historically-oriented theories of function promise to shed a great deal of light on the function of morality, and to inject new life into the centuries-old programme of drawing on human history in order to analyze (and possibly criticize) moral ideas.[29] The more we learn about early human evolution, and correspondingly about the selection pressures that were operative when moral beliefs, dispositions and practices arose, the more we can say about what they are designed to do. We can then conclude that morality has the function of solving the problems that it was designed to solve. In *The Ethical Project*, Kitcher defends just these sorts of claims.

---

[28] Kitcher, P. (1993). Function and design. *Midwest Studies in Philosophy*, *18*(1), p.380.

[29] See Hume, "Of the Origin of Justice and Property" in THN 3.2.2 and *The Natural History of Religion*, and Nietzsche, *On The Genealogy of Morals*. For a useful overview of moral genealogies, see Forster, Michael (2011) , "Genealogy". *American Dialectic* 1, no. 2, p. 230-50.

2. Kitcher and The Ethical Project

For Kitcher, "The Ethical Project" is a social-psychological phenomenon that can only exist when a group of animals possesses (1) dispositions to psychological altruism and (2) the capacity for what he calls "socially embedded normative guidance".[30] These capacities, when eventually combined with certain reactive attitudes and other dispositions (to praise, blame and punish) are, for Kitcher, definitive of human moral practice.

Crucially, his description of the evolution of human morality proceeds under the banner of a philosophical method he labels "pragmatic naturalism".[31] This approach is *pragmatist* insofar as it involves a fundamental re-orientation of philosophical attention away from questions of semantics, content and representation and towards philosophical problems that are encountered by ordinary people in the course of their lives. "The deepest impulse of pragmatism," writes Kitcher, "is to recall philosophy to an active role in human culture."[32] It is *naturalist* insofar as it resists familiar appeals not only to religious entities, but also to "Platonic forms, Aristotelian essences, non-embodied processes of Pure Reason, a priori truths… and Intuitions of the Good." In short, the pragmatic naturalist rejects the hard-and-fast distinction between philosophical activity and scientific activity, eschewing *a priori* analyses of philosophical concepts in favor of broadly scientific reconstructions of social problems and their possible solutions.[33]

One such problem is the question of how we should live. *The Ethical Project* offers an ambitious and historically detailed defense of the view that we can derive substantive answers to this question by focusing our attention on the basic function of ethics. It is here that Kitcher deploys his particular conception of function. He argues that early human social groups were beset by a "problem background", a series of social tensions generated by altruism failures. An altruism failure is "a failure to respond to the desires of another person, with respect to whom there is the potential for interaction".[34] He marshals a great deal of anthropological evidence, a few familiar game-theoretic models and some plausible psychological speculation to support the claim that our ancestors developed the ethical project in order to solve this problem. "The ethical project," he tells us, "is a social technology, one that originated against that problem background. The initial function of ethical practice was to solve the problems generated from recurrent altruism failures."[35]

---

[30] See also Sober, E. (1994), 'Did Evolution Make us Psychological Egoists?' in *From A Biological Point of View*, Cambridge: Cambridge University Press.
[31] The term does not originate with Kitcher. See, for example, S. Morris Eames (1977), *Pragmatic Naturalism: An Introduction*. Carbondale: Southern Illinois University Press.
[32] Kitcher (2013), "Pragmatic Naturalism". In Kaiser, M. I., & Seide, A. (Eds.), *Philip Kitcher: Pragmatic Naturalism*. De Gruyter, p.18.
[33] Kitcher (2013). For an extended overview of the contemporary projects that operate under pragmatist banner, see Misak, C. (Ed.). (2007). *New Pragmatists*. Oxford University Press, USA.
[34] Kitcher (2011), p. 304.
[35] Kitcher (2012), "*Précis* of *The Ethical Project*". *Analyse & Kritik* (01).

Altruism failure, for Kitcher, is the central problem that the ethical project is designed to remedy, and he argues that our own moral progress depends crucially on recognizing that this 'ur-problem' can receive new and better solutions given changing cultural circumstances. This somewhat comforting idea, however, is premised on the claim that the *current* function of ethics is to remedy altruism failure. Kitcher needs it to be the case that this is the basic function of ethics, and that it has remained unchanged since our distant evolutionary ancestors learned to live together. I will now argue that we have three kinds of reasons to think that these claims are false. My strategy in each of the sections that follow will be to: (1) identify a function of morality that Kitcher may have overlooked, and to (2) show how this creates a distinctive philosophical challenge for his view.

3. Initial Individual-Level Functions

Kitcher devotes long and impressively detailed chapters to showing that the psychological mechanisms which form the basis of ethics—psychological altruism and normative guidance—have been selected-for because they increased the reproductive fitness of *individuals*.[36] They are, in his terms, solutions to a creature's most basic problems, securing adequate food and shelter and attracting mates. If Kitcher is to remain consistently committed to his conception of function, he must say that these psychological mechanisms have the function of providing for the satisfaction of basic needs of individuals. This is an *individual-level* function of psychological altruism and normative guidance.

By contrast, when Kitcher moves to "the ethical project", he abandons the individual-level perspective and focuses on the *group*, and on the processes of cultural selection which produce ever more refined moral codes. He notes, correctly, that individual or "biological" success and cultural success can easily come apart. For example, "codes commanding obedience need not be those that further reproductive success."[37] Yet, we must still hypothesize that individuals reaped substantive benefits from the ethical project, or else the individual-level pressure to evolve resistance to that project would have overwhelmed the forces of group-selection that promoted participation in it. The original individual-level function of ethical dispositions was to satisfy individual needs, and that function must have remained operative as the ethical project got going.

This empirical result gives rise to a set of philosophical problems for Kitcher. For him, p*sychological altruism* is not merely a disposition to behave in an apparently altruistic fashion, rather, it is the propensity to make a serious effort to satisfy the needs or wishes of another

[36] On psychological altruism, he writes: "When weak animals are forced to compete for resources they need, their inability to win contests by themselves confers a selective advantage on a disposition to identify with the interests of conspecifics, particularly with those who are in a similar predicament." (p. 65) And on normative guidance, he claims that "animals with a capacity for recognizing and following orders have advantages over their fellows who lack that ability." (p. 91) While he is not explicit on this point, in these passages Kitcher makes it reasonably clear that he is referring to *individual-level* selection.

[37] Kitcher (2011), p.108

person, for no reason other than *that* they have those needs or wishes. However, on the basis of Kitcher's own evolutionary story, I can conclude that my own dispositions towards psychological altruism have a kind of individual-level function, the function of satisfying my basic needs.[38] This, again, is because the satisfaction of individual needs was the *problem* which psychological altruism evolved to solve.

First, note that this particular function is a very strange one, indeed: I must remain completely unaware of it if I am to display the disposition itself. To act in a psychologically altruistic fashion just *is* to see one's action as having the *sole* purpose of promoting the interests of another. Yet, insofar as one's action is an expression of the psychological disposition in question, this motivating thought is necessarily mistaken, since the action also has the function of satisfying one's own basic needs. A hard-headed naturalist might simply scoff at this requirement and remind us that scientific inquiry is not guaranteed to tell us that the functions of our practices or dispositions are identical or even consistent with the functions we *think* they have. But Kitcher is not just a naturalist, he is also a *pragmatist*, someone whose inquiry begins with the aim of integrating the deliverances of moral philosophy into the practical perspectives of those who must make real-life decisions. Here, the method threatens to unearth functions which cannot be integrated into our deliberation.

Now, perhaps reflection on the function of our own dispositions can take place in a "cool hour", away from the exigencies of actual decision-making. In the case of psychological altruism, however, this reflection will lead us straight to a familiar and powerful skeptical challenge. If I survey my past altruistic actions, I will almost certainly discover that many of them involved the sacrifice of certain important personal goods. If the basic biological function of psychological altruism is to allow me to acquire such goods, then I must see my dispositions as having regularly *malfunctioned.* In order to rectify this problem, I will naturally think that I ought to acquire a more finely tuned sense for when altruism will ultimately involve personal sacrifice, and I will try, in the future, to act altruistically only when I sense that there is something in it for me. In other words, reflection on the basic function of psychological altruism will lead me, somewhat paradoxically, to become what Hume called a *sensible knave*. Here, Kitcher's own account threatens to unearth functions of morality that strongly resist integration into our practical perspectives, and insofar as they can be so integrated, they form the basis for a skeptical challenge to his own, pro-social conception of morality. Whether my empirical claims about the function of altruism are ultimately correct, this remains a *kind* of problem that Kitcher does not address.

4. Initial Group-Level Functions

[38] Its function cannot be to satisfy the basic needs of anyone else, since at this level of selection, the relevant competitors just *are* all the other members of my species.

Let us now move, with Kitcher, to the cultural level, where the unit of selection is the group, and where competition between groups drives what he calls *cultural* selection. Now, many anthropologists believe that these competitions were frequently violent, and there is strong evidence that successful hunter-gatherer groups were often those that could raid, kill and defend territory more efficiently.[39] This, I claim, is an extremely important part of the problem background that confronted early hunter-gatherer groups. Moreover, while the anthropological evidence for the prevalence of inter-group conflict in hunter-gatherer societies is not decisive, there are reasons to accept that these societies *must* have engaged in serious and often violent competition.

In order to see this, consider that in outlining his preferred version of the group-selection process, Kitcher frequently cites Robert Boyd and Peter Richerson, whose model is meant to provide the general framework under which cultural competition can occur.[40] It is therefore worth pointing out that in a 1995 study, Boyd and Richerson note that their group-selection model requires, by analogy with genetic selection, the *extinction* of a significant number of competing groups. It has long been thought that this was a comparatively rare event amongst human hunter-gatherers, but Boyd and Richerson's own ethnographic research suggests that this impression is mistaken, and that group-extinctions (driven mainly by violent conflict) may have been surprisingly common. They take this to be a crucial consideration that *supports* the very model of group-selection that Kitcher endorses.[41]

Putting these considerations together, we can see that under the reasonable assumption that this not-so-rosy picture of early hominid evolution is correct, Kitcher is forced to say that at least one original function of human morality is to enable groups of humans to successfully protect themselves from, antagonize or even make war on neighboring groups. Morality has, at the group-level, two distinct original functions, to remedy intra-group altruism failure *and* to

---

[39] Azar Gat (2006, *War in Human Civilization*, ch. 1-2) shows that in the two existing areas where hunter-gatherer societies survived in relative isolation from agriculturalists and Western explorers—Australia and the Pacific Northwestern area of North America—violent warfare, raiding, the abduction of slaves and even genocide were comparatively frequent. More evidence is given in Ember, C. (1978). "Myths about hunter-gatherers." *Ethnology*, 27, 239-448; Keeley, L. H. (1996). *War before Civilization: The Myth of the Peaceful Savage*. New York: Oxford University Press; Manson, J., and Wrangham, R. W. (1991). "Intergroup aggression in chimpanzees and humans," *Current Anthropology* , 32 , 369-390.

[40] See Kitcher (2011) p. 87fn, 109fn, 112fn.

[41] Soltis, J., Boyd, R., & Richerson, P. J. (1995). "Can Group-Functional Behaviors Evolve By Cultural Group Selection?: An Empirical Test." *Current Anthropology*, *36*(3), 473-494. Elsewhere, Eliott Sober and David Wilson offer the following cautionary note:

> Our goal... is not to paint a rosy picture of universal benevolence. Group selection does provide a setting in which helping behavior directed at members of one's own group can evolve; however, it equally provides a context in which hurting individuals in other groups can be selectively advantageous. Group selection favors within-group niceness *and* between-group nastiness. (Sober, E., & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, p.9)

increase military success against rival groups.[42] Kitcher obscures this possibility by omitting any references to violent conflict in his descriptions of early human group dynamics.[43]

These considerations create a powerful type of skeptical challenge that Kitcher does not recognize and which threatens to undermine his normative ambitions. In *The Ethical Project*, he envisions and responds to an *external* challenger who wishes to overturn the original function of ethics and embrace a new way of life.[44] To my mind, a much more difficult opponent will accept that the original function(s) of ethics can, in principle, determine the basic form and content of modern ethical life. She will thus adopt a perspective *internal* to the ethical project as Kitcher is defining it. However, she will choose to emphasize the original function *I* have been outlining, that of promoting successful antagonistic competition with other groups. This challenger, who is in fact a very familiar participant in the ethical project, will correspondingly wish to articulate and defend virtues like in-group loyalty, ethnocentrism and cultural conservatism. On Kitcher's own meta-ethical view, this challenger is on very stable ground, since her conception of the good respects an original function of ethics.[45]

Now, as a matter of fact, Kitcher recommends that we view humanity as a single group, and he will likely want to argue that this challenger is clinging to a morally irrelevant distinction between her group and the rest of humanity. But this challenger will ask why Kitcher is now permitted to ignore a basic environmental feature of his own group-selection model and re-cast the moral community in terms of a single group. By contrast, her own conception of the good, which emphasizes the flourishing of her own ethnic group and the persistence of its traditions, is consistent with the problem background in which morality actually evolved, and so respects a primitive function of morality itself.

---

[42] It might be suspected that these "two" functions really are just two ways of describing the same effect. On the contrary, it *logically* need not have been the case that socially cohesive groups were culturally successful *because* they were better able to sustain (and survive) intra-group conflict. Social cohesion might simply have enabled more efficient resource acquisition and childcare, and human groups might have been more causally isolated from one another than they actually were. In this possible world, the sole group-level function of ethics might well have been to promote social cohesion. But in *our* world, it is not necessarily so: group-selection may well have been driven by actual (and not just biological) conflict.

[43] The following passage is representative: "Sometimes, they interacted with other bands, in whose practices they saw something to inspire revision of their own rules. Eventually, some groups merged, and aspects of one or both of the antecedent codes endured in the practice of the subsequent society. Some bands simply died out, or dispersed, and their ethical practices withered with them, even though survivors may have brought facets of the previous code into the groups they joined. Sometimes new arrivals, accepted perhaps as mates, brought novel ideas to the campfire discussions, producing a synthesis previously envisaged by neither of the ("parent") groups. Processes of these general types (and probably many more) combined to cause some kinds of rules to be prevalent, others rare" (2011, p. 110).

[44] (2011) p. 235.

[45] Joyce, in *The Evolution of Morality*, offers a variant of this objection: "If moral judgments are to be epistemically justified with reference to facts about human evolution, then making new and unexpected discoveries about human evolution could, and sometimes should, change our minds about enormous tracts of our moral opinion"(176). I am making essentially the same philosophical point, though I take myself to be putting something of a sharper point on it, since I am noting that it is Kitcher's own adherence to the etiological conception of function which makes it so difficult to rebut this challenge.

5. Subsequent Group-Level Functions

Finally, an example adapted from Gould and Vrba (1982) will help to illustrate the phenomenon of *functional generation*.[46] Suppose it is the case that while avian feathers were initially selected for their ability to provide temperature regulation during harsher periods of the earth's history, they *came to be* selected for their ability to aid in flight after the earth's climate became milder. Suppose it is also the case that many birds would could regulate their temperatures perfectly well if they were to become featherless, though they would certainly find it very difficult to fly (and thus to survive and reproduce). Now, what is the *current function* of avian feathers? We might say one of three things:

1. The function of avian feathers is only to regulate bodily temperature.
2. The function of avian feathers is to regulate bodily temperature and to aid in flight.
3. The function of avian feathers is only to aid in flight.

In an earlier paper, Kitcher himself discusses this example, and he concludes that (1) is not a respectable option. He notes that it is unlikely that any biologist would want to ignore the fact that avian feathers are *now* helping birds to fly. After some discussion, he adopts (3), arguing that the *present* and the *recent past* are in most cases the time-periods that are most relevant to determining the present function of an evolved trait.[47]

Returning to the history of human morality, I now want to suggest that with the invention of agriculture and the founding of large, permanent settlements, human morality began to change in ways that make serious trouble for Kitcher's view, especially given the commitment noted above. So far as I can tell, during this phase, large-scale human societies tended to have: (1) rigidly hierarchical and inegalitarian modes of social organization, (2) comparatively brutal systems of punishment, often triggered by relatively trivial crimes, (3) systems of religious belief which proscribed ritualized suffering, sacrifice and even genocide, (4) systems of slave-ownership, wherein 15-30% of the adult population was committed to forced labor, and (5) rampant militaristic campaigns against neighboring societies.[48]

Plausibly, such societies survived and propagated themselves *because* they were able to maintain these group-level features. According to a dominant theory amongst anthropologists, widespread adoption of systems of property-ownership made competition over land and resources even

[46] Gould, S. J., & Vrba, E. S. (1982). "Exaptation-a Missing Term in the Science of Form." *Paleobiology*, 4-15.
[47] Kitcher (1993), p.267.
[48] See Gat, Azar. (2008). *War in Human Civilization*. Oxford University Press; Meltzer, M. (1971). *Slavery: From the Rise of Western Civilization to the Renaissance*. Cowles Book Co; Tuchman, B. W. (2011). *A distant mirror: The calamitous 14th century*. Random House Inc; Rummel, R. J. (1996); *Death by government*. Transaction Books, Ames, C. C. (2009); *Righteous Persecution: Inquisition, Dominicans, and Christianity in the Middle Ages*. Univ of Pennsylvania Press; Pinker, S. (2011). *The Better Angels Of Our Nature: Why Violence Has Declined*. Penguin.

fiercer than during the hunter-gatherer phase, and the increased size of these societies made egalitarian or democratic modes of decision-making maladaptive.[49]

If we plug this cultural-evolutionary story into Kitcher's conception of function, we arrive at some unsettling conclusions. For even if we ignore my earlier criticisms and grant Kitcher the claim that morality *first* evolved as a response to altruism failures, it is still possible that morality subsequently came to serve new functions. It appears as though the human capacity for normative guidance and psychological altruism promoted the survival of large, post-agrarian cultural groups in virtue of the fact that such capacities enabled the creation and internalization of norms which *promoted* suffering, cruelty, inequality and war. It is difficult to see how Kitcher can avoid the conclusion that one important function of morality is to *license* or *support* a certain form of altruism failure, normally on the part of those in positions of social power.[50]

So, even if we set aside my earlier worries, we have strong reason to suspect that, like avian feathers, morality as Kitcher understands it developed entirely new functions, ones which are not necessarily consistent with or derived from an earlier function. Morality came to promote the flourishing of various societies by enabling social elites to retain and exercise their power in decidedly non-altruistic ways.[51] This might seem counterintuitive, but it is important to stress that we cannot both embrace this model of functions and then resist it when it delivers strange or even repugnant results.

Moreover, it is precisely this new function that can ground yet another familiar skeptical challenge. When Thrasymachus asserts that justice is nothing but whatever is advantageous to the strongest members of society, he is not conjuring this idea out of thin air. Rather, he is drawing on some obvious empirical facts about what is called 'justice' in various societies and about how the ruling class employs that concept for their own personal gain.[52] He might have been very pleased to discover Kitcher's meta-ethical view, according to which these functions of

---

[49] See Bowles, S. (2009). "Did Warfare Among Ancestral Hunter-Gatherers Affect the Evolution of Human Social Behaviors?". *Science*, *324*(5932), 1293-1298; Seabright, P. (2010). *The Company of Strangers: A Natural History of Economic Life (Revised Edition)*. Princeton University Press.

[50] The argument for this claim is structurally identical to Kitcher's own argument for the 'original' function of ethics. We identify a serious problem background facing large-scale societies (i.e. increased inter-group competition), and we note that the development of authoritarian, inegalitarian and militaristic moral codes was an effective solution to that problem. Of course, the mere existence of sanctioned cruelty does not establish this conclusion: morality may ultimately remedy altruism failure by sanctioning a well-regulated system of punishment directed at those whose failures threaten the survival or integrity of the group. But this is not how punishment operates during the period in question. Rather, horrific forms of punishment are devised and inflicted on those whose crimes did not threaten the social order to any significant degree. See Rummel, R. J. (1996). *Death by Government*. Transaction Books.

[51] In his critique of Kitcher, Kim Sterelny makes roughly the same suggestion. See Sterelny, Kim (2012). "Morality's Dark Past." *Anal. Kritik*, *34*(1), 95-115.

[52] Rachel Barney writes that Thrasymachus "begins like a good social scientist, claiming to discern the underlying unity behind superficially diverse phenomena: laws differ from *polis* to *polis*, depending on the nature of the regime in force, but really they are everywhere the same in serving the powers that be." Barney, Rachel, (2011) "Callicles and Thrasymachus," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2011/entries/callicles-thrasymachus/>.

ethical discourse can lead us straight to normative conclusions about how we ought to live. While I do not think any of us will want to accept the Thrasymachean outlook, the question is whether Kitcher has the resources to resist it. In light of the actual history of morality, and in light of his own professed emphasis on *recently* developed functions, it is not at all clear that he does.

So far, Kitcher's own conception of function has given voice to three characters: (1) the "sensible knave" who insists that moral behaviour ought to serve his own interests, (2) the cultural conservative who places great value on morality's ability to sustain her own group, and (3) the moral elitist, who thinks that morality ought to promote the interests of a ruling class. One effect of Kitcher's story is that it serves to portray these figures as standing *outside* of human ethical practice, as holding views that are in deep tension with the sole primary function of ethics. However, the fact that these figures are such familiar and perennial features of the normative landscape suggests that their challenges arise from "deep" aspects of human nature. It may well be the case that a dyed-in-the-wool scientific naturalist will be forced to reject Kitcher's portrayal, for these normative outlooks may well be every bit as grounded in human nature as the outlook of the pro-social, egalitarian-cosmopolitans that Kitcher hopes we will all become.

In sum, the actual history of human development provides at least three challenges for philosophers like Kitcher. First, there is the fact that morality probably has various individual-level functions which complicate the picture. Second, there is the reasonably well-confirmed hypothesis that morality developed in response to a certain kind of *competitive* selection-pressure. Third, subsequent stages of human history appear to suggest that morality may have acquired the function of supporting and maintaining certain inegalitarian social structures. All of these challenges are grounded in the same conception of function that Kitcher deploys, and in what I take to be plausible models of human history. A skeptic might wish to question the accuracy of my own claims about history, but it is important to note that even if just one of my three empirical claims is accurate, Kitcher is in trouble. One major lesson here is that historical inquiry into moral functions is not friendly to philosophers who wish to uncover a relatively simple or tidy story about the function of human morality. There are no built-in safeguards against massive functional pluralism, nor against the discovery that functions can change dramatically as time goes by, nor against the more unsettling revelation that ethics is not what it appears to be.

These doubts arise from a standpoint that is basically friendly to pragmatic naturalism, one which grants that the naturalistic functions of morality might be employed in order to generate authoritative standards of ethical conduct. I am utilizing Kitcher's own conception of function, and I am attempting to plug respectable empirical models of human development into that same conception. However, when I do so, I do not find the tranquil simplicity that Kitcher finds. Rather, I see bewildering complexity. Is there such a thing as *the* function of morality? I do not think that Kitcher has provided us with reason to think that there is. Having argued for this

critical point, I want to close on a more constructive note by outlining what I take to be a more respectable version of pragmatic-naturalist inquiry into the social functions of ethics.

6. **The Prospects for Pragmatic Naturalism**

The idea that moral philosophers might investigate the naturalistic functions of morality in human life is an exciting one. However, as I have suggested, this project faces empirical and conceptual problems of its own. Pragmatic naturalists, to my mind, should be very cautious in deploying the concept of *function*, which needs to be carefully operationalized before substantive and interesting conclusions can be drawn.[53] In particular, one question that needs to be addressed centers on *levels* of selection. As I have shown, historical evidence can reveal group-level or individual-level functions, and there are probably many more levels of which we can responsibly speak (genetic, species-wide, etc.). Which functions are we interested in, and how will we decide which among them is the most important?

Moreover, in accordance with the basically scientific spirit of their view, pragmatic naturalists must continually be on the lookout for alternative hypotheses about social functions which might complicate their own preferred theories. This involves resisting the urge to oversimplify human history, society or psychology in the service of telling a neat or unified story about ethics.

One way to take these methodological precepts seriously is to adopt the altogether reasonable hypothesis that ethics is not a single, unified thing, to accept that it is in all probability a hodgepodge of practices and dispositions that have been cobbled together in response to a myriad of shifting historical pressures. This recognition can, in turn, lead us to a productive narrowing of focus. We can take particular human dispositions or practices and try to situate their *particular* functions socially and/or historically. Indeed, such projects are already part of moral-philosophical practice: consider Joel Feinberg's classic inquiry into what he called the "expressive" function of punishment, Williams' ideas about the 'proleptic' function of blame, and Kitcher's own independent investigation into the role played by religious practices in modern societies.[54] Pragmatic naturalism need not aim to tell a unified story about morality in general, rather, it can help human individuals to understand and evaluate their own particular beliefs, dispositions or practices.

---

[53] Note that I do not claim that the concept must receive an *analysis*. As I have already claimed, and as Millikan (1989) has forcefully argued, there is no need to expect that the word "function" names some unitary concept, nor to think that we cannot fruitfully employ various conceptions of function depending on our purposes. Here, I merely ask that pragmatic naturalists articulate a defensible conception of function and that they pay close attention to what that conception entails during the course of their inquiry.

[54] Feinberg, Joel, 1965, "The Expressive Function of Punishment," *The Monist*, 49: 397–423; Williams, Bernard, (1995) "Internal reasons and the obscurity of blame." In *Making Sense Of Humanity And Other Philosophical Papers 1982-1993*. Cambridge University Press; Fricker, Miranda (ms), "The Point of Blame"; Kitcher, Philip (2011). "Militant Modern Atheism". *Journal of Applied Philosophy* 28 (1):1-13.

A second way to shore up pragmatic naturalism is to turn our focus away from speculative stories about our distant evolutionary past and focus more closely on very recent cultural history, a period which appears to have produced a spectacular array of genuinely novel moral dispositions and practices in almost rapid-fire succession. Moreover, this shift can be motivated by an analogy with biological function, which, as we have seen, is often revealed by more recent evolutionary history. This shift might also be motivated by a familiar and potentially devastating problem, one which makes trouble for my own speculative ventures as much as Kitcher's. This is the almost paralyzing paucity of data on early human societies. Available evidence radically underdetermines hypotheses about hunter-gatherer life, and Kitcher's critics might well seize on this difficulty in order to argue that it is foolish to draw *any* normative conclusions at all from what Kitcher calls "how-possibly" stories. A renewed focus on recent cultural history might serve to deflect these sorts of worries.

In sum, historically-oriented conceptions of function provide moral philosophers with a powerful tool, but this tool must be deployed with caution. Thanks to recent work in the philosophy of biology, the naturalistically-minded ethicist has a new way of bringing the history of morality to bear on various problems in moral philosophy. However, as I have repeatedly stressed, this philosopher must not seek shelter from the messy details of that same history. Perhaps one day we will achieve a deep understanding of the many functions performed by human moral practice, an understanding which will enrich and inform our practical lives. That day has not yet arrived.

Renee Bolinger
University of Southern California

Renee Bolinger is a second year graduate student at the University of Southern California. She received her masters from Northern Illinois University in 2012, and is interested in philosophy of language and social & political philosophy.

"The Specification Problem for Anderson's Democratic Egalitarianism"
Commentator: Jessica Talamantez, Northwestern University

Abstract: *The* specification problem *is a challenge to both specify the content of our egalitarian obligations, and justify these obligations on liberally neutral grounds. After briefly reviewing the content of the obligations Anderson aims to justify, I examine three strategies one may use to ground these particular obligations: appeal to an explicit principle of interpersonal justification, to an implicit principle of rejection, and to justifications based solely in equality. I argue that the first fails to validate a key inference, the second leads to a disastrous number of conflicting obligations, and the third could ground some of the right obligations but viciously over-generates. This third option could be saved by appeal to substantive normative principles, but not without sacrificing a commitment to liberal neutrality. I conclude that Anderson's democratic egalitarianism faces a dilemma:* either *it can specify the content of our obligations (which capabilities we must secure for every member of the society),* or *it can ground our obligations in neutral liberal principles, but it cannot do both.*

Democratic egalitarians contend that measuring equality in terms of distributable goods misses the point of equality. Resources and opportunities are after all only instrumentally valuable, as a means for attaining various states that Amartya Sen calls 'functionings'. A functioning may be thought of as a way of being: relevant functionings "can vary from such elementary things as being adequately nourished, being in good health, avoiding escapable morbidity and premature mortality, etc., to more complex achievements such as being happy, having self-respect, taking part in the life of the community, and so on."[55] A capability to achieve a certain functioning depends on a combination of personal and circumstantial factors. The capability to be well-nourished, for example, depends on the availability of food and one's metabolic rate: given the same quantity and quality of food, a person with low metabolism may be able to be well nourished while someone with high metabolism is unable. What ultimately matters to an individual is what capabilities she has: what sort of life she is in a position to lead, what substantive freedoms she enjoys, etc.; resources are merely means to these ends. There is an important caveat here: attaining some functionings requires making trade-offs that not every member may be willing to make. So rather than ensuring that citizens *actually* attain some set of functionings, democratic egalitarians are committed to securing capabilities: *effective access* to

[55]Sen, 1992:40.

these functionings, where an individual S has effective access to some functioning iff S "can achieve the functioning by deploying means already at her disposal."[56]

The Specification Problem

Shifting the focus in this way from generic goods to capabilities results in what I'll refer to as the SPECIFICATION PROBLEM: granting that capabilities are the relevant metric of equality, what obligations do members of a just state have to each other? If we interpret the *egalitarian* aspect of democratic egalitarianism as requiring that all citizens be ensured the minimal capabilities for functioning as societal or moral equals, then it induces a sufficientarian standard, and requires us to identify the relevant set of capabilities which we are obligated to guarantee to each citizen. The specification problem then becomes the problem of finding satisfactory answers to two questions:

> THE CAPABILITIES QUESTION:
> What minimal set of capabilities do we have an obligation to ensure?
>
> THE GROUNDS QUESTION:
> What grounds our obligation to ensure *these* capabilities, rather than others?

These two questions highlight a tension between two desiderata: (i) to give an answer to the first question that respects our intuitions about which capabilities must be secured, (ii) to supply principles that are neutral between reasonable conceptions of the good that generate a set of required capabilities matching our answer the first question.

Answering the Problem: A Search for Perfection

Advocates of various forms of democratic egalitarianism have answered the specification problem in different ways. Amartya Sen, Martha Nussbaum, and Elizabeth Anderson provide a good variety of responses, representing the full range of strategies. Sen's solution is to require actual democratic selection: on his view, each society must select for themselves which capabilities will be in the guaranteed set.[57] So the answer to the GROUNDS question is *because society agreed on this set*, but the price of this strategy is that the answer to the CAPABILITIES question must be left up entirely to the negotiations of the society. Nussbaum chooses the other horn, answering the CAPABILITIES question with an extensive list of the capabilities she takes to be required for human flourishing.[58] Many of these capabilities come into conflict with various religious and cultural values, so in order to secure this particular set Nussbaum must invoke "an

[56] Anderson, 'What is the Point of Equality?' 1999:318.

[57] A. Sen, *Development As Freedom* 1999: 74-75.

[58] M. Nussbaum, *Women and Human Development: The Capabilities Approach* 2000:78-80.

objective normative account of human functioning"[59] as an answer to the GROUNDS question, sacrificing liberal neutrality.[60]

Anderson aims to steer a middle path: she answers the CAPABILITIES question with a partially specified list of particular functionings, and aims to satisfy GROUNDS by presenting liberally acceptable principles that will generate a set including (at least) all of the specified capabilities. I argue that Anderson's egalitarianism faces a dilemma in answering the specification problem: either she cannot give a principled reason why we should be obligated to secure just her preferred set of capabilities, or she must appeal to a substantive conception of the good, securing her desired set by sacrificing her commitment to liberal neutrality.

Which Capacities?

Anderson's official formulation of the necessary set of capabilities is highly unspecific. She writes, "negatively, people are entitled to whatever capabilities are necessary to enable them to avoid or escape entanglement in oppressive social relationships. Positively, they are entitled to the capabilities necessary for functioning as an equal citizen in a democratic state."[61] With this established as the broad guiding principle, Anderson circles back to give some more illuminating examples.

Positive Entitlements

To function as a democratic equal, an individual must have all the capabilities relevant to functioning (i) as a rational agent, (ii) as a participant in the system of cooperative production, and (iii) in the civic government. Participation in civic government requires some obvious capabilities–voting, engaging in political speech, petitioning the government—and some less obvious ones, like access to public spaces & services, being accepted by others, and having 'equal standing.'[62] The capabilities necessary for participating in cooperative production include access to the means of production, education to develop one's talents, free occupational choice, the rights to make contracts, receive fair pay for labor, and have one's contribution recognized by others.[63] Importantly, Anderson allows that an individual's access to these capabilities may be made conditional on exercising her other capabilities: receipt of a fair wage may be ensured *conditional on* performing labor in a job, so long as the individual has access to such a job and

[59]M. Nussbaum, 'Nature, Function, and Capability: Aristotle on Political Distribution' 1988. In her more recent work, Nussbaum has changed her position to closely resemble Sen's: the list is a "thin vague conception of the good," and all the details must be settled by each society in practice. (Nussbaum,'Love, Literature, and Human Universals: Comments on the Papers', *Martha C. Nussbaum: Ethics and Political Philosophy* 2000:137-139.)

[60]Severine Deneulin contends that even Nussbaum's revised position fails liberal neutrality in 'Perfectionism, Paternalism and Liberalism in Sen & Nussbaum's Capability Approach' (2002).

[61]Anderson 1999:316.

[62]Anderson 1999:317-8. To have equal standing is roughly to have an equal ability to make claims, register complaints, and be owed justification for policies affecting you.

[63]Anderson 1999:315.

can perform the labor. The capabilities required for human agency involve knowledge of one's circumstances and options, the ability to deliberate, and 'the psychological conditions of autonomy.'

Negative Entitlements

The negative entitlement amounts to the real ability to exit hierarchies of three types: (i) standing, (ii) esteem, and (iii) command, and this entitlement is insensitive to facts about how the individual came to be in such hierarchical relationships. To secure exit from hierarchies of *esteem* the state must (a) avoid making any official moralizing judgments concerning projects or reasonable conceptions of the good, and (b) foster a plurality of views of the good, with the aim of securing a set of conceptions such that no group or project is held in contempt by all competing views.[64] A hierarchy of *command* occurs when one person is subject to the will of another due to the asymmetry of their bargaining positions. Some such hierarchies are necessary for the organization and production of a public good (and even these must be subject to strict controls),[65] but any unnecessary command hierarchies pose a threat to the weaker party's autonomy. Though citizens should be free to enter into such relationships, they are entitled to the social means of exit. This capability of exit from relationships of command is closely conceptually linked to the positive entitlement to the 'conditions of autonomy,' and while both are amorphous as listed, the intended content becomes clearer in Anderson's discussion of their application to the case of dependent caretakers.

The Case of Dependent Caretakers

Individuals who take primary caretaking responsibilities (for dependent children or elderly relations) devote a substantial portion of their time and energy to discharging these duties. This displaces wage labor they might otherwise have performed, and often results in financial dependence on a primary wage earner. This dependence in turn develops into vulnerability: the caretaker faces high barriers to exit, since s/he does not have access to a living wage in the absence of the wage-earner. Anderson contends that in such situations the caretaker is entitled to the capability of exit: in this case, that means that we have an obligation to provide him/her with access to a living wage conditional on work consistent with the performance of their caretaking duties.

These then are the entitlements that constitute Anderson's answer to the CAPABILITIES question; it remains to be seen whether she is able to provide a principle that simultaneously justifies

[64] Anderson is insistent on this point. She writes, "The expected and preferred outcome of such liberty is a plurality of conceptions of the good, which generate rival and cross-cutting orders of esteem, such that no group comes out on the top or bottom of everyone's rankings, ... and no esteem raking counts as official, as one to which everyone is expected to defer." (Market Risks, 264; precisely the same wording recurs in 'Expanding the Egalitarian toolbox: Equality and Bureaucracy', 145.) This entails that the state may not enact policies that (explicitly or implicitly) rank some conceptions of the good as less valuable than others.

[65] Some such constraints include that "hierarchies must be justifiable to those subject to command, where the terms of justification secure their free consent – not as a matter of desperation, given a lack of reasonable alternatives." (Market Risks, 264).

obligations to secure these capabilities and does not extend to cover a wide range of capabilities that are intuitively not obligations of justice.

On What Grounds?

Anderson's Explicit Principle

Anderson's explicit principle is grounded in her concept of what is involved in recognizing each other as equals. Acknowledging one another as moral equals requires that "each accepts the obligation to justify their actions by principles acceptable to the other and ... take mutual consultation, reciprocation, and recognition for granted."[66] In addition to recognizing each other as moral equals, members of an egalitarian society must acknowledge each other as participants in a system of cooperative production, with the consequence that "workers and consumers regard themselves as collectively commissioning everyone else to perform their chosen role in the economy."[67] Equality therefore mandates acceptance of Cohen's principle of *interpersonal justification*, IJ:

> IJ: Any consideration offered as a reason for a policy must serve to justify that policy when uttered by anyone to anyone else who participates in the economy as a worker or a consumer.[68]

IJ implies that a policy is permissible only if there is at least one ϕ such that ϕ passes the interpersonal justification test and is a consideration justifying the policy.[69] The IJ as given is a purely negative constraint on the sorts of arguments that may be offered in favor of a policy, and says nothing about the necessary or sufficient conditions for a policy to be obligatory. Presumably if there is no argument that could justify a policy α, then (and only then) it is obligatory to refrain from αing.

Inadequacy & the Fix

To show that IJ grounds the target obligations, Anderson notes that in discharging duties of care, dependent caretakers provide a social good, freeing the rest of us to devote our time to the

---

[66] Anderson 1999:313.

[67] Anderson 1999:322.

[68] Cohen, 'Incentives, Inequaliity, and Community' (1991):280.

[69] Cohen's gloss on the test is actually quite helpful:

> ...an argument for a policy satisfies the requirement of a justificatory community, with respect to the people it mentions, only if it passes the interpersonal test. And if all arguments for the policy fail that test, then the policy itself evinces lack of justificatory community, whatever else may be said in its favor.
> Now, an argument fails the interpersonal test, and is therefore inconsistent with the community, if relevant agents *could* not justify the behavior the argument ascribes to them. ... The thesis associated with the interpersonal test is that, if a policy justification fails it, then anyone proposing that justification in effect represents the people it mentions as *pro tanto* out of community with one another. ... The interpersonal test focuses on an utterance of an argument, but what it tests, through examination of that utterance, is the argument itself.'

production of market goods. So, caretakers serve a role in the scheme of cooperative production, and we should consider their actions as collectively commissioned by the other participants. We could therefore describe the policy of failing to secure caretakers' access to a living wage as acting according to the principle:

P1. Let us assign others to discharge our caretaking obligations to dependents, and attach such meager benefits to performance in this role that these caretakers live at our mercy.[70]

Anderson contends that such a principle cannot plausibly pass the interpersonal justification test, and concludes from this fact that "dependent caretakers are entitled to enough of a share of their partner's income that they are not vulnerable," or, if this is not sufficient, justifies "socializing some of the costs of dependent care."[71] It isn't immediately obvious how an obligation to provide follows from the failure of P1. Probably the thought is something like this: P1 is a permission claim, equivalent to

P1*. It is permissible to assign others to discharge our caretaking obligations to dependents, and attach such meager benefits to performance in this role that these caretakers live at our mercy.

when a permission claim to $\neg\alpha$ is false, the corresponding obligation to α is true; so if P1* fails, then it is not permitted to fail to compensate caretakers, ergo it is obligatory to provide for them. This would all follow if the object of the test, P1, were an *action* or *policy*, and the test a test of the permissibility of actions or policies directly. However, the test only operates at the level of principles, and all that is shown by the failure of a principle is that it cannot be used to justify the target policy. Absent proof that there is no other principle that could justify α, the failure of ϕ yields information only about (im)permissible justifications, not policies. We could get the target obligations from the failure of P1 by adopting a test like IJ*:

IJ*: If a principle fails the interpersonal justification test, then we are obligated to refrain from acting according to the policies it purported to justify.

The Disaster: Inconsistent Obligations

There is good reason to be wary of a test like IJ*. Principles, after all, are justifications for policies, and the same policy may be justified in a variety of ways and at various levels of detail. If we accept IJ*, then for any policy α, if any one of the principles potentionally justifying α fails interpersonal justification, then it follows that we have an obligation to $\neg\alpha$. Given that bad principles can be generated to attempt to justify nearly any policy, this is a recipe for generating conflicting obligations.

[70] Anderson 1999:324.

[71] Anderson 1999:324.

Two of the ways that such a conflict could arise are of particular interest; the first is by allowing principles to vary in generality. Suppose that γ is a policy, the performance of which entails performing two more specific policies, α and β. Consider two principles, ω (which justifies $\neg\gamma$) and ϕ (which justifies the policy β): if ω fails, then by IJ* we have an obligation to perform γ, which entails performing α and β. But if ϕ fails too, then that generates an obligation to perform $\neg\beta$.[72] The second source of conflict occurs at a single level of generality. Consider a pair of principles μ and ν, such that μ justifies an action α and ν justifies an omission $\neg\alpha$. Both principles may fail interpersonal justification, generating obligations to $\neg\alpha$ and α, respectively.

These conflicts result from a more general problem in the structure of IJ*. To validate the inference from the failure of P1 to an obligation to provide for dependent caretakers, the test must entail that if someone can reasonably object to a principle that justifies refraining from αing, then we have an obligation to α. But in a reasonably pluralist society, there will be many policies α such that someone may object to α, and someone else may object to $\neg\alpha$. In fact, this will hold for any policy that does not enjoy a consensus. A principle like IJ* is unacceptable for a society committed to liberal neutrality: rather than permitting reasonable disagreement over policies that lack overlapping consensus, it would systematically ground contradictory obligations to act and refrain from acting according to the policy.

Implicit Principles?

We have seen that attempting to use IJ* to validate Anderson's argument for our obligations to caretakers leads to disaster, but this alone does not show that the principle she explicitly endorses, IJ, fails to justify the target obligation set. It may be that we can still get IJ to ground the right set of obligations by invoking different background assumptions; perhaps that only reasons grounded in claims of fundamental equality will be acceptable as spoken by anyone to anyone else.[73]

---

[72]To make the case more concrete, you may note that the LIMITLESS case developed in the next section follows this pattern. Refusing to provide the means of freedom from the command of the dealer without requiring an addict to give up the habit would count as an instance of m, while providing such means is an instance of n:

> m -- Let us fail to treat the members of society as equals, securing for them unequal civil capabilities, including freedom of conscience and freedom from dominance.
>
> n -- Let us bear the cost of providing greater resources to those who, knowingly and intentionally, for the sake of a short-term benefit that others chose to forego, made it the case that it is more costly to secure for them freedom from dominance.

The failure of m grounds an obligation to provide unconditional means, while the failure of n grounds the counter-obligation to refrain from making such provisions.

[73]There's some reason to think that Anderson intends this sort of assumption to augment IJ; she remarks that

> "I prefer to base the case for [the distributive arrangements'] legitimacy on the ground that they secure everyone's entitlements to the material conditions of their freedom and equality on a basis of reciprocity, with everyone interacting with one another on terms all can accept." (Anderson, Market Risks, 253-4.)

How far can we get by assuming that (a) only reasons based in equality will survive interpersonal justification, and (b) these reasons must make essential reference to reciprocity and securing the material conditions of freedom? Is it possible to get Anderson's target result in the dependent caretaker's case? Suppose we grant Anderson's claim that some relationships of command threaten individual's freedom, and that the remedy for such a threat is "social arrangements that secure each adult's personal independence."[74] Then if we can show that the dependent caretaker is in the sort of relationship that threatens her freedom, we can rely on our assumptions to derive the dual conclusion that (i) the proposition that we ought to provide for such caretakers *will* survive interpersonal justification, and (ii) there will be no acceptable reason to refrain from making such provisions.

So the question now is *in virtue of what does the dependent caretaker's situation warrant our concern*? It's certainly not that she is literally *unable* to exit the relationship. Nor is it that she is suffering from some unchosen bad luck: the caretaker's vulnerability results from a complex of factors, including an intentional initial decision to have and/or care for children.[75] Rather, the relevant fact is that her financial dependence, given her duties, makes the cost of exit excessively high. To leave, she must either accept a minimal/unacceptable standard of living or give up the performance of her duties–a project that she takes to be of central importance.

Abstracting from the details of the particular case, the principle that grounds our obligation to dependent caretakers appears to be something like P3:

P3: An agent S is entitled to the social means of exit if S is in a situation such that

1. There is some project or good ϕ that imposes high costs on S,
2. ϕ is subjectively very important to S,
3. Commitment to ϕ makes S vulnerable to asymmetrical relations of command.

Working backward, if we claim that every situation that fits this scheme constitutes a threat to the individual's freedom, then we can use our earlier assumptions to ground an obligation to provide the means of exit to dependent caretakers.

Over-generation

So far, so good, but since nothing we've said so far restricts possible values for ϕ, we appear to have offered a principle that grounds not only provision of exit for dependent caretakers, but also a variety of far less intuitive cases, for example, the LIMITLESS case:

> LIMITLESS: Suppose that Phil has developed an addiction to a designer drug D. He knew the risks of addiction before he began taking D, but heard it had a great high, and might give him a competitive edge in the market. At first, this was true: Phil's creativity

[74] Anderson, Equality and Bureaucracy, 146.

[75] Anderson is adamant that democratic egalitarianism is not a starting gate theory, and does not withhold aid based on judgments of responsibility: "Some outcomes are so bad that they are objectionable even if they are the consequence of a voluntary choice." (Market Risks, 257).

skyrocketed, allowing him to design intriguing trinkets. Over time, his body developed a tolerance for D and it ceased to have creativity-enhancing effects, but Phil was still able to manufacture and sell his trinkets successfully. Eventually, the D dealer started asking Phil to pay $x more for D (where x is some amount beyond what Phil could reasonably afford), or else do him certain favors. Phil could escape the dealer's influence by quitting the drug, but Phil is unwilling to do this because that would trigger withdrawal symptoms he considers unacceptable. Consequently, unless Phil is provided with the means of affording to pay his dealer $x, he is subject to the will of the dealer.

This case meets all the criteria: (a) continuing to take D imposes high costs on Phil, (b) D is subjectively very important to Phil, and (c) this commitment makes Phil subject to an asymmetrical relation of command. Nevertheless, it is intuitively wrong to say that we have an obligation to provide Phil with the means to support his D addiction. To conclude from this that P3 *viciously* over-generates, we must satisfy ourselves that Anderson would want to exclude this case from the scope of legitimate entitlements. We may convince ourselves of this by reviewing Anderson's critique of Van Parijs' proposal to provide all citizens with access to as high a basic income as is sustainable:

> "The chief difficulty with his proposal is that his basic income would be awarded to all unconditionally, regardless of whether they were able or performing socially useful work."[76]

Her criticism can be read as having two motivations:

1. Unless it incorporates some requirement that recipients perform socially useful work if able, the proposal fails to give others reason to bear the cost of supplying this income, and
2. Guaranteeing such an indiscriminate basic income will result in the sustainable level being much lower than it would be if made conditional on the performance (if able) of socially useful work.

Providing Phil with $x appears liable to both aspects of this critique: the reason that Phil requires support is a personal project with minimal social benefit; this provides no obvious justification for others to bear the costs of the project, and any income supplementation we offer to Phil reduces the amount of capital we have available to provide for others. This gives us at least a *prima facie* reason to think that if P3 justifies providing for Phil, it grounds an obligation that Anderson denies we have. Unless we can find some way of restricting its scope, P3 appears as unable to ground obligation to secure all and only the capabilities Anderson identifies in her answer to the CAPABILITIES question.

Playing Defense

It is illegitimate to attempt to solve the case by offering Phil a deal: we'll provide a one-time provision of $x, on the condition that Phil goes to rehab. In LIMITLESS, Phil's continuing to take

[76] Anderson 1999:299.

D is the project ϕ that corresponds to a caretaker's duties, so until we have established a principled disanalogy between the cases, such an offer is equivalent to offering caretakers a living wage *conditional on* giving up their caretaking duties. Since this move doesn't qualify as securing the means of exit from relationships of command in the caretaker case, to count it as doing so here would encode an official judgment that Phil's project is not worthwhile, which is inconsistent with the State's obligation to avoid making any such official value judgments concerning projects or life-plans. Similar remarks apply to offers to provide assistance conditional on gradual reduction in Phil's D intake: if continuing to take D is of great importance to Phil, then we cannot justify making our aid conditional on this sort of (even gradual) sacrifice without licensing a parallel move in the caretakers case, or else making an official judgment that Phil's projects are less valuable.

We might instead try adding the requirement that ϕ be a socially useful project. This stipulation would still get the right result in the CARETAKERS case, since the duties performed are socially useful, and would potentially block purely recreational drug habits, but is of no avail in blocking the LIMITLESS case. This is because, as developed, taking D has some positive effect on Phil's market-related activities, and Anderson's threshold for 'socially useful work' is low–low enough to count activities as useful if they have useful effects. Raising the threshold to require direct involvement would get the wrong result in the caretakers case, since their work is only indirectly-market affecting. Changing the measure of 'socially useful' from market-related activity to something involving a more substantive notion of *useful* or *good* could potentially separate the cases, but would require an official judgment that Phil's project is strictly less valuable than the caretakers', and making such an official judgment would violate the conditions necessary for *equality of esteem* as well as sacrificing liberal neutrality. So it looks like this option is a non-starter. One approach that is sure to separate LIMITLESS from the dependent caretakers is to stipulate that ϕ must be central or important for a good life. If this condition is to distinguish caretaker cases from drug addiction cases, then the criterion of importance cannot be determined by what *the agent* finds centrally important; any ϕ that satisfies condition (b) will be subjectively central/important (at least to S) for a good life. So if this condition is intended to narrow the field, it must be by appeal to some sort of *objective* measure of worthwhile projects or goods. The catch, of course, is that unless this can be specified in some neutral way (and I do not see how it could be), appealing to such a list violates liberal neutrality.

So, we cannot save P3 by stipulating that ϕ be socially useful or of central importance to a good life. Absent any other means of restricting the values for ϕ, P3 looks doomed to over-generate, and thus is not an acceptable grounds for the obligations that Anderson takes us to have.

Conclusion

We have tested three strategies for generating Anderson's targeted list of capabilities: appeal to an explicit principle of interpersonal justification, to an implicit principle of rejection, and to

justifications based solely in equality. The first failed to validate a key inference, the second succeeded in that but led to a disastrous number of conflicting obligations and violated liberal neutrality, and the third could ground some of the right obligations but viciously over-generated. The final option could be saved by appeal to substantive normative principles, but not without sacrificing a commitment to liberal neutrality. I conclude that Anderson's democratic egalitarianism faces a dilemma: *either* it can specify the content of our obligations (which capabilities we must secure for every member of the society), *or* it can ground our obligations in neutral liberal principles, but it cannot do both.

REFERENCES

Anderson, Elizabeth. 'What is the Point of Equality?' *Ethics*, Vol. 109, No. 2 (January 1999), pp. 287-337.

_____. 'Expanding the Egalitarian Toolbox: Equality and Bureaucracy', *Proceedings of the Aristotelian Society* 2008: suppl. vol.

_____. 'How Should Egalitarians Cope with Market Risks?' *Theoretical Inquiries in Law* Volume 9, Issue 1, 2007:239–270

_____. Justifying the Capabilities Approach to Justice', in Robeyns & Brighouse (eds.) *Measuring Justice*, Cambridge University Press, 2010:81-100.

Deneulin, Severine. 'Perfectionism, Paternalism and Liberalism in Sen & Nussbaum's Capability Approach.' *Review of Political Economy* 14, no. 4 (10, 2002): 497-518.

Cohen. 'Equality of What? On Welfare, Goods, & Capabilities', in Nussbaum & Sen (eds.) *The Quality of Life*. Oxford: Clarendon Press, 1993.

Nussbaum, Martha. *Women and Human Development: The Capabilities Approach*. Cambridge University Press, 2000.

_____. 'Non-Relative Virtues: An Aristotelian Approach', in Nussbaum & Sen (eds.) *The Quality of Life*. Oxford: Clarendon Press, 1993.

_____. 'Nature, Function, and Capability: Aristotle on Political Distribution', *Oxford Studies in Ancient Philosophy,* 1988:suppl. vol.

Pogge, Thomas. 'Can the Capabilities Approach be Justified?' in Robeyns & Brighouse (eds.) *Measuring Justice*, Cambridge University Press, 2010: 17-60.

Sen, Amartya. *Development As Freedom*. Oxford University Press, 1999.

_____. 'Capability and Well-being' in Nussbaum & Sen (eds.) *The Quality of Life*. Oxford: Clarendon Press (1993):30-53.

_____. *Inequality Reexamined*. Cambridge, Mass: Harvard University Press, 1992.

Abraham Roth
The Ohio State University

Abe Roth is Associate Professor in the Philosophy Department at Ohio State University. He works mainly in the philosophy of action, with a recent focus on shared agency. Papers include "Shared Agency and Contralateral Commitments" (*Philosophical Review*, 2004). "Prediction, Authority, and Entitlement in Shared Activity" is forthcoming in *Noûs* and stems from a previous talk given at NUSTEP.

"What is it to Accept a Promise?"
Commentator: Stephen White, Northwestern University

Abstract: *The* acceptance *of a promise bears on the normative significance of the promise itself; promissee acceptance arguably is a necessary condition for the obligation to keep one's promise. But what is it to accept a promise? I suggest a new account in terms of intentions: for B to accept A's promise to φ is for B to intend A's φ–ing. The thought is that the distinctive role of intentions in practical reasoning will help us to understand the agency of the promisee in accepting a promise. I then turn to Cognitivism about intentions, the view that one's intention involves a distinctive, non-evidentially warranted belief or expectation regarding the φ-ing intended. Cognitivism and the account of promissory acceptance are then used to defend Scanlon's expectation-based view of promissory obligation against recent criticism by Kolodny and Wallace.*

The *acceptance* of a promise is normatively significant; it is, arguably, a necessary condition for the obligation one takes on in promising. But what is it to accept a promise? I suggest a new account in terms of intentions: for B to accept A's promise to φ is (among other things) for B to intend A's φ–ing.[77] This approach, conjoined with Cognitivism about intentions (the view that one's intention entails the belief regarding the φ-ing intended), is then used to defend Scanlon's expectation-based view of promissory obligation. The approach to promising taken here makes use of some of the resources from the study of shared agency. I close by critiquing another view in the same vein. But let me begin with why acceptance is necessary for promissory obligation.

It is a good thing to be able to give or receive assurance that some action will be performed, and one way this is done is through promising. In his discussion of promising, Scanlon rightly emphasizes the importance to us of this value of assurance, of being able on particular occasions to take for granted what others will do when one could not otherwise form any reliable expectations about their actions.[78]

But sometimes I don't want assurance, such as when you promise something burdensome or irritating for me. If we keep in mind that promising is supposed to benefit the promisee – to provide assurance or to demonstrate that one takes seriously something that is important to the

[77] This will need to be qualified to address cases raised by Kit Fine.
[78] Scanlon, "Promises and Practices", *Philosophy and Public Affairs*. See also his *What We Owe to Each Other,* Ch. 7.

promisee, it becomes less plausible to think that the promising generates obligation when it is unwelcome. Why think that the resources of moral motivation, obligation, blame, sanction, etc. associated with promising can be harnessed in a way that works against the promisee?

So it's likely that *some condition* would prevent obligation being generated in the cases of unwelcome promising. What condition would that be? A minimal suggestion is to require merely that the promisee *want* the promised φ–ing.[79] But sometimes a promisee is of two minds on the matter: I might want the puppy because it's so cute, but also want not to have it because of the responsibility. Or the want in question is preliminary, tentative, or fleeting. It's not obvious in these cases that a promise that doesn't receive a reply, that is not *accepted*, generates an obligation. This verdict is supported by the thought that if the promisee wants what's promised but hasn't fully decided and delays replying, she cannot assume that the promise is in force when she finally settles in favor of it; she would have to contact the promisor to see if he is still willing to take up the obligation. This suggests that there is no obligation until the promise is accepted.[80]

If promissory obligation requires not merely some want on the part of the promisee, but also their acceptance, then it would seem that the promisee has more control over whether the obligation is in place. Why might this be a good thing? One normatively distinctive feature of promising is that the promisee is in a special position to release the promisor from the obligation.[81] This feature imposes some responsibility upon the promisee to exercise this power of release when appropriate, and sometimes the promisee might not want to take up this responsibility. Another aspect of promising is that when the promisor acts, he is arguably acting at least in part on behalf of the promisee, fulfilling a promise to her.[82] But even if the action is of the sort that is welcomed by the promisee (e.g. the promisor might be donating resources and time to some cause the promisee cares about), the promisee might not like the idea that the promisor is doing it *for her*. She simply doesn't want to be *involved* with him to that extent, or perhaps to any extent at all. (She might think, "If he wants to spend his time and money on that, he should just do it for himself and not because of some promise to me!") So if it is correct to think that promising involves forging a special relationship between individuals,[83] there are reasons for thinking that it cannot be undertaken unilaterally by the promisor; the promisee

---

[79] In order to rule out threat-promises, Scanlon invokes a sophisticated condition in this vein. He requires that the promisor knows the promisee wants to be assured about the φ–ing (*ibid* 216, 218).

[80] This speaks against Shiffrin's suggestion ("Promising, Intimate Relationships, and Conventionalism", *Philosophical Review* 2008) that we can make do without the acceptance condition; on her view, the relevant condition is just that the promisee not reject the promise. There are further reasons to think that the absence of rejection is insufficient to generate promissory obligation. Sometimes a promise amounts to an inappropriate overture to which the promisee would not deign a response. Presumably there is no obligation to keep such a promise. Or consider the possibility that one might be overwhelmed with promises and unable to keep track of which to reject. (It's hard enough to keep up with email.) If, as I hope to show, it matters whether one can control which promissory relationships one enters into, then such a case suggests that one must positively accept the promise in order to generate the obligation. (I think that implicit acceptance is also possible, as when a promise is solicited by the promisee. But implicit acceptance is not merely the absence of rejection.)

[81] Scanlon builds this into his Principle of Fidelity, which underlies promissory obligation. This aspect of promising is central to what Shiffrin calls the *rights transfer view*. ("Immoral, Conflicting, and Redundant Promises", in Wallace, Kumar, and Freeman, eds., *Reasons and Recognition – Essays in Honor of T.M. Scanlon*), with references to Raz, Shiffrin, and Owens.

[82] This requires some elaboration and defense. Something like it is held by Shiffrin in "Immoral, Conflicting, and Redundant Promises."

[83] Darwall reference.

would have to play a role in establishing the relationship – presumably by *accepting* the promise. So, considerations of autonomy – of having a say or some control over one's interpersonal relationships – also speak in favor of an acceptance condition for promissory obligation.

We might arrive at the acceptance condition along a different line of thought, focusing on (for want of better terminology) epistemic considerations confronting the promisor. If a promise is solicited by the promisee, we may presume that this is not a case of an unwelcome promise. But not all cases are so simple. The promisee's attitudes can be a complicated nexus of wants, many of which are contextualized and relativized. Take a case where by making a promise the person generates a preference in the promisee, one that doesn't exist independently of the promise. For example, you and I have been living in different parts of the country for some time. We often talk of seeing each other but, to your consternation, my trip plans on previous occasions have fallen through. Now, however, I promise to be out in your area in May. Being that we're old friends, you have some obligation to see me on my visit. You do want to see me, and you are glad that I'm not visiting in April or June, when you are swamped with deadlines and other commitments. But, as far as you're concerned, you most prefer that I visit later in the summer – a time that doesn't work well for me. Having made the promissory overture, am I obligated to visit? If we only look to promisee wants, it's hard to know what to say about the case. You want me to visit in May more than in June. But, independently of my making the promise, you want me to visit later in the summer, and it's only because I promised that you acquire some preference for me to visit in May. Is this want good enough to generate the obligation? How can I tell?

These concerns disappear if you were to *accept* the promise. Acceptance lays bare the want and relevant background circumstances, clarifying whether something desirable in the abstract really is wanted in the circumstances at hand, outweighing other considerations. An acceptance condition would therefore seem to solve a sort of epistemic problem that the promisor faces. The acceptance cuts through all the questions we have about preferences and conditions, etc., and settles the issue. An account of promissory obligation that doesn't avail itself of the notion of acceptance would run the risk of being much harder to apply correctly.[84]

The considerations outlined suggest including amongst the conditions for promissory obligation that the promisee accept the promise. Such a condition can be found in Thomson's account of *word giving,* which encompasses promising: "Y gives X his or her word that a proposition is true only if Y asserts that proposition to X, and (i) in so doing, Y is inviting X to rely on its truth, and (ii) X receives and accepts the invitation (there is uptake)."[85] Thompson incorporates acceptance as a condition on promising itself. But we could instead have acceptance as a condition on promissory obligation, allowing for the possibility that some promises are not taken up and don't generate promissory obligations.

If acceptance is crucial for generating promissory obligation, we might wonder just what it is to accept a promise. What does Thomson have in mind when she says that there must be *uptake*? One suggestion is that acceptance is simply coming to believe or expect that the

[84] This assumes that the promisee has access to their own preferences, and we might wonder about this. But I take it that the epistemic problem for the promisor is solved at least in the sense that it's no longer the promisor's problem to solve. The epistemic burden is shifted so that it resides with the promisee now; and that's a problem the promisee has anyway, just from being an agent.
[85] Thomson, *The Realm of Rights,* 298.

promisor is going to do as he says he will. But this doesn't seem right. Suppose that I eavesdrop on or otherwise overhear A promising B that he'll do X and, given what I know about A's reliability and how seriously he takes his commitments, I come to believe that A will do X. There is some sort of uptake here, but surely this isn't to accept the promise, since I am after all not in a position to do so (I cannot, for example, release A from the obligation). Perhaps the proposal about acceptance in terms of belief leaves out a positive attitude, evaluation, or acknowledgement of what's on offer. But similar considerations from eavesdropping (as well as other arguments) show that such pro-attitudes will not do the trick.[86]

Thomson is not explicit about why acceptance is needed, nor does she elaborate on the nature of acceptance. But she associates promising with an *invitation* to rely on what the speaker is promising, and so is thinking of acceptance as a reply to an invitation. This suggests an element of control on the promisee's part – something along the lines of choice or some other exercise of the will, which fits with the considerations of autonomy outlined above. Although Thomson never says this, I want to interpret her talk of uptake as involving an intention, or something very much like it. That is, I suggest that B's acceptance of A's promise to φ involves B intending A's φ–ing.[87]

There are a number of reasons for understanding promissory acceptance in terms of intention in this way, of which I mention two. First, if all goes well in promising, the promisee B receives assurance about the promisor A φ–ing. The matter of φ–ing is settled, whereas it very well may have been up in the air beforehand. Notice, further, that the *promisee* has settled the matter. After all, the assurance received is presumably a matter of the promissory obligation incurred by the promisor and the seriousness with which the promisor takes the obligation.[88] But if I am correct in thinking that there is no obligation until there is acceptance, then not only is the matter settled, it is settled by the promisee. With this in mind, consider the now-standard view of intentions, *viz,* that it is precisely the sort of attitude the having of which involves some practical or deliberative question being settled.[89] So if accepting the promise to φ settles the matter of φ–ing, then acceptance is playing the role that we expect intention to play. This is reason to think that it amounts to intending the φ–ing.

The second reason to think this is that the acceptance of promises seems to be subject to a consistency requirement. Suppose A promises to meet me for lunch today, and that B promises to complete a report for me by 3pm today. Suppose also that A meeting me for lunch is not compatible with B finishing the report (for example, B cannot finish the report without getting A to skip lunch and help him on it). Then I can't legitimately accept both of these promises. One of them must be rejected. But this makes acceptance similar to intending, for intentions, too, are subject to a consistency requirement.[90] The demand for consistency in the promises one accepts (not to mention the demand for consistency between one's intentions and the promises one

[86] More to say here.
[87] This is not intended as a sufficient condition for acceptance. Discuss Fine's worry.
[88] "This time I won't be late; I *promise*." (Shiffrin ref.) See Scanlon *ibid*; Kolodny and Wallace "Promises and Practices Revisited" *Philosophy and Public Affairs,* 2003 make this clear.
[89] Harman, *Change in View;* Bratman, *Intention, Plans, and Practical Reason*.
[90] Harman, Bratman references.

accepts) would receive a straightforward explanation if we think of the acceptance of promises on the model of intending what's promised.[91]

The thesis that the acceptance of a promise involves the promisee intending the promisor's action may be somewhat surprising. After all, we don't usually speak of one person intending another's actions. Or if we do, it's usually thought to be morally, if not conceptually, problematic. So it might strike one as implausible that such a thing would be *required* for promissory obligation. Why not think of acceptance in more modest terms, as the promisee intending the promisor to be obligated to him?

I grant that accepting a promise entails at least implicitly intending to be in a promissory relationship with the promisor. But there are several reasons for thinking that this is not enough to count as acceptance. First, if one agrees with Scanlon that the fundamental role of promising has to do with the value of assurance, then recall what it is that we want to be assured about. We want assurance about something that they will do. It's the φ-ing that we want to count on, not merely that someone is obligated to do it. So, what one is *accepting* is not limited merely to being in a relationship where the promisor is obligated to φ. If we only receive assurance that the promisor is obligated, then that is a sign that something has gone wrong and that promising in this instance is not doing what it's supposed to do. So if we agree with Scanlon about this role of promising, and we want (as I do) to understand acceptance in terms of intention, then we should hold that the promisee, in accepting the promise, is willing or intending more than that the promisor be obligated. She's intending the promisor's φ-ing.[92]

The consistency constraint on accepting promises also speaks in favor of this conclusion. If, in accepting a promise, the promisee merely intends the promissory relationship – the promisor being obligated to her – then it's not clear why the promisee cannot accept the incompatible promises described above – where A promises to meet me for lunch, and B promises to complete the report but can only do so by working through lunch with A. After all, even if it is impossible for both A and B to meet their obligations to me, my *intention* for A to be obligated to me is entirely compatible with my *intention* for B to be obligated to me. (Imagine that my goal is to see one or the other fail in meeting their obligation to me.) The consistency constraint on acceptance therefore cannot be explained in terms of the consistency constraint on intentions if the promisee is only intending the promisor to be obligated. In contrast, we do account for the consistency constraint on acceptance if what the promisee intends encompasses the actions promised. My intention for A to meet me for lunch and my intention for B to deliver the report are incompatible, given the facts about the world; the corresponding acceptances would then also be incompatible.

The explanatory benefits of thinking that acceptance involves intending the promisor's actions might be out of our reach if we cannot get over our reservations about the idea of one person intending the actions of another. I won't attempt to fully address this concern here. But let me say a couple of things to try to make this claim more palatable. First, I should make clear

[91] Further considerations for thinking that acceptance should be understood in terms of intention…

[92] Although assurance is fundamental to promising, that doesn't mean that we all instances of promising involve assurance. Given this, someone might object along the following lines: A intending B's φ–ing might suggest that A cares about B's φ–ing. But this seems to suggest that one cannot accept promises that one doesn't care about. I think that this objection reads too much into intention. We often intend things that aren't so important to us; we don't care so much one way or another, but we intend just to get the matter resolved.

that in defending the thesis that acceptance involves the promisee intending the promisor's action, I don't mean to be suggesting that the promisee guides or controls the promisor's action in the way that an agent guides or controls her own movements at the time of action. That sort of control is exercised by the promisor as he carries out what he's promised; the promisee is not controlling the promisor as if the latter were some marionette. In what sense, then, does the promisee intend the promisor's action? The proposal is that the promisee's acceptance settles the practical matter much in the way that one's prior intentions and decisions can settle what one will later do. Thus, when it is time for the promisor to act to fulfill the promise, he doesn't reopen the question of whether to φ. Instead, he just guides his action in a way that takes for granted what was settled for him or her to do when the promisee accepted.[93]

So, we get a better idea of the sense in which one person intends the action of another by focusing on the role of intentions in settling deliberative questions. But this now gets to the second worry: if B is settling what A does, then doesn't this threaten to undermine A's agency in a morally problematic way? I concede that settling what another does can sound awfully like commanding them. But we should keep in mind that B only has this power because A made a promise. Moreover A's promise wasn't that he'd do *anything* that B told him to. As a result, B's power is strictly limited (he can only settle or secure A's φ-ing).[94]

I turn now to Scanlon's expectation-based view of promissory obligation. My aim here is not to present anything like a full defense of it. I only want to use the proposal about the acceptance of promises to address one sort of criticism that has been directed against Scanlon's view.

According to Scanlon, underlying the obligation to do as one has promised is the following "Principle of Fidelity":

> **Principle F:** If (1) A voluntarily and intentionally leads B to expect that A will do X (unless B consents to A's not doing so); (2) A knows that B wants to be assured of this; (3) A acts with the aim of providing this assurance, and has good reason to believe that he or she has done so; (4) B knows that A has the beliefs and intentions just described; (5) A intends for B to know this, and knows that B does know it; and (6) B knows that A has

[93] But is it really settled? Scanlon's case of the Profligate Pal, and Shiffrin's point about the possibility of education or reform are concerns that need to be addressed. The worry in both is that the promisee has no expectation regarding whether the promisor will follow through, and yet the promisee can accept. It seems that in such cases, the promisee cannot genuinely accept promise, and that there is no promissory obligation here. Which is not to say that there are other forms of commitment here, e.g. that which comes with a vow, or the sorts of commitment associated with shared activity.

[94] One way to understand how B has this power is in terms of the rights-transfer view of Shiffrin and others (check Shiffrin on characterization). Just as one might, simply by so willing, consent to allow another a (moral) right to certain forms of intimacy, so one might, simply by so willing, promise and thereby allow another some limited form of morally legitimate authority over one's actions. This transfer of a right over how the promisor will act is then understood as the moral obligation the promisor has to follow through on a promise. But the rights-transfer view needn't be the only way for B to have this power. I might instead simply put myself qua promisor into a situation wherein the promisee is able to issue a limited form of intention for me and thereby settle what I will do. No special moral authority need enter the picture just yet; just as whatever authority one's own intention might have in one's practical reasoning and action at this stage is at best a form of rational normativity, and doesn't necessarily amount to a *moral* obligation to do as one intends. The *moral* obligation to follow through on the promise and do what the promisee accepted (i.e. do what the promisee intends for the promisor to do) could then be understood to stem from the expectation-based moral principles of the sort defended by Scanlon. See below.

this knowledge and intent; then, in the absence of special justification, A must do X unless B consents to X's not being done. (Scanlon, *What We Owe One Another,* 304)

On Scanlon's view, we do not need to appeal to a convention or practice of promising in order to explain what is wrong with not keeping one's promise. Of course, if there is a practice or convention involved in promising and not keeping one's promise would undermine or free-ride on this useful practice, then this would be a *further* reason for the promise-breaking being wrong. But, according to Scanlon, this is not the fundamental story of why it's wrong to break a promise. The obligation, and the wrong in failing to meet it, are understood in terms of the Principle F, which on Scanlon's Contractualism is a principle regulating action that no one could reasonably reject.

Scanlon himself raises a circularity or "Can't Get Going" worry for his account. As noted at the outset, a fundamental purpose of promising is to provide assurance. By subjecting oneself to the obligation that comes with making a promise to X, the promisor acquires a substantive reason to perform X; this assures the promisee, who can now expect X to be done. (Scanlon 306-7, 322.) On Scanlon's view, however, the promisee's expectation is a condition for the promisor being obligated. So it's hard to see how the promissory obligation so understood can fulfill its role in generating an expectation in the promisee.[95]

My proposal that we think of the acceptance of a promise in terms of an intention can help Scanlon here. To see this, we need to turn briefly to the philosophy of action. An important claim for my purposes here is that intention involves the belief or expectation regarding the φ-ing intended. This can be seen as an extension of the related thesis that intentionally φ-ing entails at least believing, if not knowing, that one is φ-ing.[96] One could hardly be said to be φ-ing intentionally if one doesn't even realize that one is φ-ing. Likewise, it would be odd to say that one intends to φ without any expectation that one will.[97] More theoretical considerations also speak in favor of Cognitivism. It offers a way to relate normative constraints on intentions (e.g. consistency and means-end coherence mentioned earlier in the paper) with similar rational constraints on beliefs.[98] And it suggests why intentions play a role in planning so that what's intended can figure in further practical and theoretical reasoning, not only by the agent but by

[95] **[Very long footnote recounting Scanlon's solution and why Kolodny & Wallace are not satisfied…]**

[96] The stronger knowledge claim is if typically attributed to Anscombe, *Intention*, 11, as well as Hampshire & Hart, "Decision, Intention, and Certainty", *Mind* 67, 1958. I've followed e.g. Setiya ("Practical Knowledge", *Ethics* 118 (2008)) in citing Anscombe here as a source for Cognitivism. But given what she says about direction of fit, it is not entirely clear that Anscombe's view amounts to the Cognitivist claim about *believing* that the action intended will be performed.

[97] See Hampshire and Hart, 5-6; Grice, "Intention and Uncertainty", *Proceedings of the British Academy 57,* 1971, 278.

[98] Harman, "Practical Reasoning" *Review of Metaphysics* 29:3, (1976), Velleman "What Good Is a Will?" Anton Leist & Holger Baumann (eds.), *Action in Context*. de Gruyter/Mouton (2007); Setiya, "Cognitivism about Instrumental Reason" *Ethics* 117 (2007). This is not to say that the norms of intention can be reduced to those of belief. See Bratman, "Intention and Means-End Reasoning", *Philosophical Review;* "Intention, Belief, and Instrumental Reasoning" David Sobel and Steven Wall, eds., *Reasons for Action* (Cambridge: Cambridge University Press)

others as well: if in intending to go shopping, I expect that I will, then I can count on the fact that there will be food available for dinner.[99]

I think that a good case can be made for Cognitivism, but my focus here will be on its implications for promising. As a preliminary, note that Cognitivism has an interesting – some would say troubling – epistemic implication: if I intend to φ, the corresponding belief or expectation that I will φ doesn't seem to be formed on the basis of any evidence. According to the Cognitivist, the expectation that I will φ is simply a part of, or else just comes with, the intention to φ. The intention, however, is not formed on the basis of evidence. When, for example, I decide and intend to do some grocery shopping, it's not normally because the evidence indicates that this is what I'm up to. It would be absurd to think along the following lines: *I am walking down High Street; I don't usually walk here unless I'm on the way to Big Food-n-Deal Supermarket and I'd only go there if I'm going to shop for groceries; so it must be that I'm intending to do some shopping*. Rather, the intention that I will do some shopping is immediate upon the decision to go shopping, and normally none of this is inferred from the evidence.[100] So if the expectation automatically comes with the intention, then it will normally be evidentially ungrounded.

No doubt, this is a reason why some hesitate in embracing Cognitivism. One might try to water down the Cognitivist claim, suggesting instead that we think of the expectation that comes with intending as something less than full out belief that one will be φ-ing. Maybe it's the belief that it's more likely than not that one will be φ-ing, or perhaps that at least it's more likely than if one weren't to intend. But notice that even the weaker belief conditions raise much the same epistemic worries about ungroundedness. And if we try to make do without *any* belief component whatsoever, then it becomes hard to see what the point is supposed to be of making a decision and intending some end.

I will not here tackle the issues that arise in the epistemology of self-attribution of intentions; no doubt things get even trickier on the Cognitivist conception of intention.[101] But if the Cognitivist is right about intentions, then, unless we embrace a skepticism about a significant portion of what is traditionally included as part of our agency, there must be an important non-evidential component in the warrant for the expectation regarding the intended φ-ing.[102]

Let's turn back to promising. We now have the resources to address the Can't Get Going worry. I've argued that in accepting the promise, the promisee forms an intention regarding the promisor's action. If Cognitivism is right, then the promisee has a non-evidentially warranted

[99] Grice, 272; Harman. The planning role of intention is most fully developed by Bratman, although he rejects cognitivism. Ironically, in insisting on intention being fundamental, Bratman might be closer to Anscombe than many who see themselves as inspired by her. Another prominent critic of Cognitivism is Davidson.

[100] Shoemaker, Moran references. This is not to say that empirical and evidential considerations have no role to play. Beliefs about what I can do, or what sorts of skills I could possibly acquire, or what I might be prompted to do in different circumstances, etc. certainly figure in the background of my acquisition of the belief that I will φ and what it is that I think I can intend. And something might come up to undermine or defeat intention-based expectations regarding my φ-ing. (Casullo, Burge references.) But such empirical considerations by themselves are not enough normally to account for how I arrive at the intention. (Remark about Moore's Paradox statements.)

[101] Cognitivists sometimes appeal the expectation being self-fulfilling in order defend its epistemic legitimacy. See Harman, Velleman. For criticism, see Langton. (refs.)

[102] Again, this non-evidential component can be supplemented with empiricial background knowledge regarding what sorts of things one can do, and the circumstances that one is in.

expectation regarding what was promised. (There may be a background evidential component so that one believes that one is in a position to intend the action of the promisor; this no doubt involves some understanding of the promisor's moral competence and conscientiousness so that if they were morally obligated to X, then they would X.) The expectation then triggers Principle F. So, thinking of acceptance in terms of promisee intention not only gives a more intimate and active role for the promisee in generating promissory obligation; it also assimilates the ungrounded expectation needed to trigger Principle F to a form of expectation that typically is not based on evidence.

In presenting their criticism of Scanlon, Kolodny and Wallace consider a response that can be seen as sharing an element of what I'm proposing here.[103] According to what they call the "power of positive thinking" response, the promisee goes ahead and forms the expectation that the promisor will perform, knowing that once that expectation is formed, it will be justified because Principle F will be in force and the promisor (because of his moral conscientiousness) will be motivated to come through. K&W reject this response, noting (1) that it is not under promisee's voluntary control to form the belief that the promisor will do as promised without any prior evidence for it, and (2) that such a groundless belief in any case would not satisfy the conditions of Principle F.

Let me address the (2) first. K&W think (correctly) that F requires the promisor to *lead* the promisee to form the expectation. But they go on to claim that a groundless belief, since it is not based on evidence, could not amount to the promisee being led to form the belief. That's why K&W say that the power of positive thinking does not satisfy the condition of the promisor leading the promisee "to believe, on the basis of evidence," that the promisor will perform the act. (K&W, 133). But Principle F doesn't include any condition that the promisee must be led *on the basis of evidence*. So we address (2) if we can point to a way in which the promisor can lead the promisee to form a belief without the presentation or manipulation of evidence.

But how is this possible? Well, one way to do it – and this is to address (1) – is to invite the promisee to settle whether the promisor will do the relevant X. By offering a sufficiently attractive invitation (which most promises are), the promisor could be said to lead the promisee to accept the promise, even though the acceptance is voluntary. And the promisee accepts the promisor's invitation by voluntarily intending the promisor's action. This addresses the concern that the promisee expectation is not based on evidence.[104] For, if Cognitivism is true, the expectation that comes with the intention to X is not normally thought to be based on evidence. So even if there is a worry about the ungroundedness of the expectation, it is an issue raised more generally by the distinctive non-evidential epistemology associated with agency. The evidential groundlessness of the expectation upon which the Can't Get Started Problem trades, therefore should not disqualify Scanlon's understanding of promissory obligation.[105]

---

[103] Kolodny and Wallace, "Promises and Practices Revisited" *Philosophy & Public Affairs*, 31.2, 2003. They credit M. Pratt, "Scanlon on Promising," *Canadian Journal of Law and Jurisprudence* 14 (2001): 143-54.

[104] There is, of course, an evidential component in that one thinks that the promisor takes his moral obligations seriously, such that *if* he were obligated by F, then he'd perform the action. Without this, promising would not be a way to offer assurance.

[105] Critical remarks concerning rival "shared agency" accounts of promissory obligation…

Jennifer Lockhart
Auburn University

Jennifer Lockhart is an assistant professor of philosophy at Auburn University. Her research interests include Kierkegaard's method of indirect communication, Kant's practical philosophy, and issues that arise at the intersection of the philosophy of sex, love, marriage, and gender.

"The Necessity of Duty and the Open Texture of Morality"
Commentator: Heidi Giannini, Wake Forest University

Abstract: *This paper argues that the conception of duty ordinarily employed by Kantians makes it impossible to account for the more flexible features of our moral lives, the "open texture of morality." One resource that the Kantian should be able to draw upon in theorizing the open texture of morality is Kant's account of imperfect duties. I argue, however, that given the standard conception of duty, Kantians are, correspondingly, unable to provide a coherent analysis of imperfect duties. This paper argues for a revised conception of duty that harmonizes with the open texture of our moral lives and makes conceptual space for imperfect duties. On the view advanced here, our duties should not be understood as those actions that we cannot avoid performing. Rather, duties designate those actions that are worth doing, irrespective of our contingent ends.*

1 Introduction

Duty is a central concept for Kantian ethics. To be moved to act by duty is to act through a recognition of what reason says we *must* do. As Kant says, "Duty is the necessity of an action from respect for law."[106] But what does mean to say that an action *must* be performed or that an action is *necessary*?

The fact that Kant analyzes the concept of duty in terms of necessity has suggested to many that such seemingly exceptionless commands as, "Repay your debts!" or "Do not give false witness!" are, for the Kantian, paradigmatic instances of the requirements of duty. It is true that placing these kinds of imperatives at the center of a moral framework allows for a relatively straightforward understanding of the way in which the actions required by duty are necessary: rationally speaking, the actions are not optional, so that we have no choice but to perform these actions insofar as we are guided by reason. Nevertheless, this conception of morality provides a less than inspiring vision of our moral lives. On this conception, reason seems to stand over us like a strict schoolmaster, barking out orders from which we are not allowed to deviate.

This conception of morality—one that demands behavior that is not rationally optional—surely is an ill-fitting analysis of the concerns and motivations that govern the reality of our day-to-day moral lives. No doubt, some cases arise in which morality requires us to do something that is not optional. On occasion we are called like Luther to say, "Here I stand. I can do no

[106] Immanuel Kant, *Groundwork of the Metaphysics of Morals*, 4:400. All translations are from *Practical Philosophy*, ed. Mary J. Gregor (Cambridge: Cambridge University Press, 1999). All translations are given in the notes by the volume and page number of *Kant's Gesammelte Schriften*, ed. Royal Prussian (later German) Academy of Sciences (Berlin: Georg Reimer, later Walter deGruyter & Co., 1900-).

other." But many genuinely moral concerns and motivations do not reduce us to just one option: We help this person rather than that one (although had we helped that one it could have been morally motivated as well); we instantiate intimacy and respect in this form of relating to one another rather than some other (although had we related in a different way that could have equally been a way of manifesting respect); we balance a plurality of goods against each other, all of which make some moral claim. The sort of Kantian who analyzes duty in terms of non-optional actions faces difficulty in accounting for these more flexible features of our moral life, what I will refer to as the "open texture of morality."

These reflections are not meant simply to constitute an *external* critique of Kantian morality. Officially, the Kantian has resources that she can tap to account for the open texture of morality: the notion of imperfect duties suggests that moral action can be, in some way, optional. This means, however, that there is also a problem *internal* to Kantianism: How can the Kantian make intelligible the notion of an imperfect duty? How can the Kantian reconcile the emphasis on latitude in fulfilling imperfect duties with the notion of duty as "the necessity of an action from respect for law"?[107] If we have latitude in performing some particular action, in what sense is the action necessary?

In this paper I have one narrower and one broader aim. My narrower aim is to demonstrate that there is indeed a problem with imperfect duties that Kantians have not been able to solve, and the way out of this bind is to give the proper account of what it means to say that our duties are necessary. I will argue that the way to avoid the problem of imperfect duties is to embrace a conception of duty according to which *actions can be at once necessary and rationally optional.*

My broader aim is to draw out the implications of this solution to the problem of imperfect duties for identifying one of the core features of a morality centered on duty and necessary action. The important thing about practically necessary actions should not be, I will argue, that regardless of what we want, (rationally speaking) we have no other option but to perform such actions. Instead, the point of the Kantian's emphasis on duty should be that regardless of what else we may want, we have reason to perform such actions, even if there remains a plurality of other actions that we have equal reason to perform. The point of Kantian duty, I will argue, should not be that there are certain things that we cannot avoid doing. Rather, it should be that there are certain things that are worth doing, irrespective of our contingent ends.

2 The Problem of Imperfect Duties

The conception of duty ordinarily employed by Kantians is the **Rationally Non-Optional Conception (RNC)** of duty, according to which an action fulfills a duty only if, rationally speaking, the action is not optional. RNC unpacks the necessity of duty in terms of the non-optionality of the actions that are required by the moral law. RNC poses Kantians with a prima facie problem: how to make sense of what Kant sometimes calls imperfect duties. These are duties such as the duty of beneficence—the duty "to promote according to one's means the happiness of others in need, without hoping for something in return."[108] Although there are disputes concerning how to understand imperfect duties, Kant is clear that these duties allow a certain degree of latitude in how we are to fulfill them. Kant writes that the law, "leaves a

[107] Ibid., 4:400.
[108] Kant, *The Metaphysics of Morals* 6:453.

playroom (*latitudo*) for free choice in following (complying with) the law, that is, that the law cannot specify precisely in what way one is to act and how much one is to do by the action for an end that is also a duty."[109]

Given that there is latitude in fulfilling imperfect duties, reason does not dictate exactly which actions one is to perform in fulfilling these duties. Therefore, one sometimes fulfills such a duty by performing an action that is rationally optional. According to RNC, this should not possible, since duty is understood precisely in terms of the non-optional nature of the actions that fulfill a duty. Therefore, RNC has no way to account for the latitude involved in fulfilling imperfect duties. I will call this *the problem of imperfect duties.*

In order to illustrate this problem, let us consider an example. Suppose that I am trying to figure out how to spend my Saturday afternoon. I am considering helping a friend, who is moving, to pack his truck. Although my presence would be a help to him, others have already volunteered, so that if I do not help, I will not be leaving him in the lurch. I am also thinking of going to the gym in order to promote my health and to develop my physical capacities. We could put flesh on the bones of this example in many ways, but given what I have said so far, the following conclusions about this example are plausible: Although the moral law requires me to adopt the end of promoting the happiness of others, the latitude that I am afforded in complying with the law means that I am not required in this particular case to promote the happiness of my friend.[110] Helping my friend to pack his truck, therefore, is rationally optional. I would be rationally justified in helping him, but I would also be rationally justified in going to the gym.

Now, suppose I decide to help my friend move. I do it because I have been moved by reason to adopt the happiness of others as my own end, and I recognize this as a way to realize that end. Helping my friend to move is, therefore, a matter of fulfilling an imperfect duty and my action is one performed from the motive of duty.

Given the analysis I have offered of this example, RNC conflicts with the conclusion that in helping my friend to move I am fulfilling an imperfect duty. According to RNC, an action fulfills a duty only if rationally speaking the action is non-optional. Helping my friend to move, according to the above analysis of the example, is rationally optional. Therefore, helping my friend to move cannot, given RNC, fulfill a duty.

I will now consider two lines of response open to the defender of RNC. The first is the **Rigorist's**. The Rigorist attempts to avoid the problem by claiming that the actions that fulfill imperfect duties are not in fact rationally optional. The second strategy is the **Teleologist's** response to the problem. The Teleologist tries to avoid the problem by claiming that we have a duty only to adopt certain obligatory ends, not to undertake particular actions in pursuit of those ends.

The first way to try to salvage RNC is to interpret the discussion of latitude very narrowly. According to the Rigorist, although the moral law cannot specify in advance what is required in order to comply with it, in particular situations where, say, beneficence is relevant,

[109] Ibid., 6:390.
[110] I am, after all, also required by the moral law to adopt the end of developing my physical capacities.

reason will specify exactly one action that is rationally required.[111] This action is, therefore, according to the Rigorist, rationally non-optional. With respect to the example I considered above, the Rigorist would reject my analysis of the situation according to which there is more than one rationally viable alternative. The Rigorist must interpret the example differently, making the case that reason specifies whether I am to pack the truck or to go to the gym. Of course, it is not enough for the Rigorist to reject the analysis I offered of this particular case, which was intended only to illustrate a general problem. The Rigorist owes us an argument that his own analysis could, in principle, plausibly apply to every instance involving a question of fulfilling an imperfect duty.

The Rigorist manages to avoid the problem of imperfect duties at the cost of assimilating imperfect duties to perfect ones. The Rigorist holds out little promise for fruitfully analyzing the open texture of morality. The Rigorist's version of Kantianism finds itself open to the charge of being exceedingly inflexible and rigoristic.

A second response to the problem of imperfect duties comes in the form of the Teleologist's claim that one has a duty only to adopt certain obligatory ends, not to perform particular actions in pursuit of those ends. Marcia Baron takes this approach. She argues that beneficent actions are not "(under most circumstance) morally required. . . . It is morally required that one help others, but not here and now."[112] Baron claims that, "because acts of helping another are not, as individual acts, generally morally required, it does not make sense to speak of someone performing them from duty."[113]

Let us return to the present example. The Teleogist can admit that reason need not specify whether I am to help my friend pack or to go the gym. But the Teleologist must claim that whichever of these I choose, I am not fulfilling a duty in helping to pack or in exercising. The only duty involved, according to the Teleologist, is to adopt the obligatory ends that either of these particular actions would be in pursuit of. The Teleolgist would deny the second part of my analysis: that when I help my friend pack, I can act from the motive of duty and can thereby fulfill an imperfect duty.

Officially, the Teleologist professes a great deal of latitude in fulfilling imperfect duties. According to her, we are afforded free choice with respect to when, how and to what extent we pursue obligatory ends such as the happiness of others. However, the Teleologist shares the problem with the Rigorist that she is unable to account for the open texture of our moral lives, although the way in which the Teleologist fails to do so is less obvious. The latitude involved in fulfilling imperfect duties is not, according to the Teleologist, *moral* latitude. Free choice enters the Teleologist's picture at the level of actions that are situated squarely outside the realm of duty and morality. Duty, we could say, runs out after the adoption of an obligatory end. One is afforded latitude in the particular actions one chooses in pursuit of this end, but this is a purely instrumental latitude. The Teleologist cannot offer us a view according to which an open texture

[111] Although the Rigorist's position is one that is worth discussing for the sake of conceptual clarity, I know of no Kantians who explicitly endorse a view as rigoristic as the Rigorist's. For an overview of the interpretive debates surrounding the latitude involved in imperfect duties see Melissa Seymour, "Duties of Love and Kant's Doctrine of Obligatory Ends" (PhD diss., 2007, Indiana University), Chapter 4.
[112] Marcia Baron, 'Overdetermined Actions and Imperfect Duties,' in *Moralische Motivation: Kant und die Alternativen (*Hamburg: F. Meiner, 2006), 27.
[113] *Ibid.*, 28.

is woven out of our genuinely moral pursuits and concerns when those involve, say, the particular actions undertaken to promote the happiness of some particular person. The Teleologist appears flexible when it comes to the question of whether we should perform particular beneficent actions, but this flexibility comes at the cost of failing to appreciate the way in which *these particular actions* can themselves be paradigm instances of moral concern in which one is moved to act from duty.

3 The Necessary Reason Conception of Duty

My aim in this section is to solve the problem of imperfect duties by rejecting RNC. I will argue that an action can be at once necessary and rationally optional. If we can make sense of actions that are necessary while at the same time rationally optional, then we can account for both the way in which an action that fulfills an imperfect duty is necessitated by the moral law and the way in which latitude is afforded the agent with respect to whether or not she performs such an action. This solution to the problem of imperfect duties will require us to revise our conception of duty. I offer an alternative conception of duty, the **Necessary Reason Conception of Duty** (NRC), according to which an action fulfills a duty only if it is an instance of a type of action all of whose instances are good. This means that in every situation in which an agent is contemplating an action of this type, the agent has a reason to perform the action and no decisive reason to do otherwise.

My argument is as follows:

(1) An action is necessary if it is an instance of a type of action all of whose instances are good.
(2) An action that is an instance of a type of action all of whose instances are good can be rationally optional.
(3) Actions that fulfill imperfect duties are instances of a type of action all of whose instances are good.
(4) Actions that fulfill imperfect duties are necessary actions. (1 and 3)
(5) Actions that fulfill imperfect duties can be rationally optional (2 and 3)
(6) Actions that fulfill imperfect duties are necessary actions and can be rationally optional. (4 and 5)

(1) is my proposal for a better understanding of the necessity involved in Kantian duty. My argument for (1) lies in its fruitfulness. If (1) succeeds in allowing us to solve the problem of imperfect duties and offers a richer notion of Kantian duty, this speaks strongly in its favor.

In order to see why (2) is a plausible claim, we need to be clearer on what it means to say that an action is good. Consider two possibilities for what it means to say that an action is good. First, one might mean that an action is good if one has reason to perform that particular action *rather than any other available action*. This would mean that a good action is one that stands over and above all the other possibilities available as the only rational option. But this is an exceedingly strong criterion for good action. A second possibility for how to understand good action seems more plausible: an action is good if there is reason to perform that action, and there is no decisive reason to do otherwise (where this means that doing otherwise is not the only rational option). According to this second, weaker, notion of good action, in the absence of a decisive reason against performing an action, simply having a reason that speaks in favor of

performing an action is enough to make that action good. To say that an action is good is to say that there is reason to perform the action and that one would not be mistaken to perform it.

Good actions, then, can be rationally optional. In some situations, there may be one good action that stands out above all the others, so that performing that action is the only rational option. It may be tempting to think that this is always the case—that there is always some reason, however small, to perform one action rather than any other. This, however, is false. Buridan's ass type cases can be constructed in which it is stipulated that two options exist neither of which is rationally superior to the other. In these sorts of cases there are two good actions a person could perform. Performing either one of the actions in such cases would be good, but rationally optional.

One might assume that a good but rationally optional action is a conceptual possibility that can be stipulated but is seldom or never realized in our lives. In fact, the opposite is true. In our ordinary lives, we are presented with many good options, and we perform some of these actions because we see that there is some good in doing them (and no decisive reason not to)—not because we conclude that the only thing rationality permits is performing the particular actions we choose. Very seldom do we conclude that were we not to act as exactly as we do, we would open ourselves to rational criticism. This conclusion would require taxing and abstruse reasoning that would itself be highly impractical and in most cases would prove impossible.

That many of our rationally justified actions do not bear the credentials of "the only rational option to pursue" does not point to a failure of rationality on our part. It is not a matter of sloppiness or of neglecting to deliberate hard enough. Rather, that the actions that we undertake with good reason are not the only rational option for us is a consequence of the fact that practical reasoning itself, like morality, exhibits an open texture.

Openness enters into practical reasoning in at least three ways. First, there are instrumental means to our ends that are, for our purposes, *rationally commensurate* with one another. The main thing is that we take one of these good means to our ends. Any particular means from among the commensurate ones is rationally supported, but also rationally optional. In the messy cases we find in ordinary life, the point need not be that (as in an ideally constructed case) the options we face are entirely commensurate in every respect. The point can be simply that we don't care about the ways in which the options are different—for us they are practically commensurate. Second, if we admit that there is a plurality of ends some of which are *incommensurable* with one another, openness enters into practical reasoning in another way. I may weigh up the goods involved in a monastic life and the goods involved in a political life. It is at least plausible to think that there is not a single metric according to which these goods can be compared to one another in ways that must yield an answer about which life is rationally superior for me. Again, I may choose between furthering the happiness of one friend and furthering that of another. It is plausible to think that the happiness of the two friends is not a matter of goods that can be put on a single scale and weighed against each other in quantitative terms. This means that there is little reason to suppose that rationality will, in each case, dictate that I help one friend rather than another. The lack of quantitative commensurability of the ends of the friends' happiness means that this holds true even in cases where I can help one friend to a much greater extent than the other. Thirdly, constitutive means to our ends can manifest both sorts of openness depending on the case: constitutive means may be practically *commensurate* with one another or they may be *incommensurable* with one another.

We are now in a position to see that what I have been calling the "open texture" of our moral lives is a reflection of the open texture of practical reason itself—practical reason need not deliver up just one rational course of action. The three types of openness that I have been discussing all derive from the way in which practical reasoning deals with ends—how to effect or constitute those ends and how to weigh disparate ends against one another. For the Kantian, imperfect duties lie at the intersection of morality and the pursuit of ends. Imperfect duties arise when ethics introduces "the system of the *ends* of pure practical reason."[114] It is fitting that the open texture of morality should mirror the open texture of practical reason when it comes to imperfect duties.

What I have argued so far is that good actions can be rationally optional. For exactly the same reasons, an action that is an instance of a type of action all of whose instances are good can be rationally optional (premise (2) above).

Now let's turn the third and final premise: Actions that fulfill imperfect duties are instances of a type of action all of whose instances are good. This will be true if the actions that the Kantian generally holds to fulfill imperfect duties (e.g. beneficent actions) are in every instance good.

It is not implausible to think that benevolent actions are good in every instance. Here I will briefly consider two objections to this claim. The first is that some benevolent actions, e.g. stealing from the rich to give to the poor, are morally impermissible. I wish to deny that such actions (actions that are in fact morally impermissible) ought to be considered "benevolent actions." It must be part of the notion of a benevolent action that it involves promoting the happiness of others in ways that are compatible with the moral law.

The second objection is that situations arise in which there is decisive reason against performing a benevolent action because one could benefit oneself or another to a great extent by not performing the benevolent action. In order to answer this objection, let me return to a point made earlier—that Kantians need not be committed to the view that each person's happiness is quantitatively commensurate with the happiness of others. If, in attempting to weigh one person's happiness against the happiness of another, I am comparing incommensurable ends, then it becomes harder to motivate the claim that situations arise in which there is decisive reason against performing a particular benevolent action.

4 Conclusion

NRC avoids the problem of imperfect duties by allowing for the possibility that the actions that fulfill imperfect duties are both necessary and rationally optional. The necessity of an action that fulfills a duty is a matter of the action's being necessarily rational, not of the action's being rationally non-optional. NRC demonstrates, therefore, that a duty-based morality can be in harmony with the open texture of our moral lives. Recognizing a duty-based morality that can accommodate openness requires us to re-envision the role duty ought to play in moral theory. The important thing about the concept of duty should not be that there are certain things that we cannot avoid doing. Rather, the concept of duty indicates that there are certain sorts of actions that are, in every instance, worth doing, irrespective of our contingent ends.

[114] Kant, *The Metaphysics of Morals*, 6:381.

Mavis Biss
Loyola University, Maryland

Mavis Biss completed her PhD at the University of Wisconsin-Madison in 2011 and is Assistant Professor of Philosophy at Loyola University Maryland. She specializes in moral philosophy, with particular focus on Kant and Kantian ethics and conceptions of moral creativity. Her dissertation, *Moral Imagination in an Ethics of Principle*, was supported by the American Association of University Women and she has authored articles in *History of Philosophy Quarterly*, *Hypatia*, *Southern Journal of Philosophy* and *Philosophy Compass*.

"A Kantian Response to the Problem of Reception"
Commentator: Daniel Groll, Carleton College

Abstract: *Cheshire Calhoun has posed a compelling and overlooked challenge to theorists interested in moral innovation, arguing that in "ill-formed social orders" acting well according to an objective moral standard can actually produce a significant form of moral failure. Specifically, in societies characterized by multiple systems of oppression, acting rightly may require acting in ways that will be unintelligible to many people with whom one interacts. If failures of reception of right action undermine attempts at moral expression, one may "do the right thing" and yet meet moral failure.*

I draw on recent work in Kantian ethics that acknowledges relational aspects of autonomy and the conditions of socially embedded rational agency in order to resist Calhoun's conclusion. I will argue that we can appreciate the "deep sociality" of morality without supposing that cooperating in a shared system of meaning is itself a moral ideal that can come apart from the ideal of doing right. The fact that failures of moral communication may be stages in communicative processes that extend far beyond an initial series of actions and responses supports the conclusion that moral innovators' misread actions are best understood as incomplete moral successes.

Recently several feminist philosophers have drawn attention to the ways in which individual agents may imaginatively modify conventionalized moral understandings to conceptualize their experience more accurately and reconceive their possibilities for moral action. Cheshire Calhoun has posed a compelling and rather overlooked challenge to theorists interested in such moral innovation, arguing that in "ill-formed social orders" acting well according to an objective moral standard can actually produce a significant form of moral failure.[115] Specifically, in societies characterized by multiple systems of oppression, acting rightly may require acting in ways that will be unintelligible to many people with whom one interacts. The woman who openly embraces her lesbian identity and is read as exhibitionist rather than as self-respecting serves as a paradigm example of the problem Calhoun has in mind. Did she succeed or fail in expressing self-respect? What about a woman who, in trying to avoid misplaced gratitude, is perceived as rude? Because successful expression requires others' comprehension, Calhoun argues, we may "do the right thing" and yet meet moral failure.

[115] Cheshire Calhoun. "Moral Failure," in *On Feminist Ethics and Politics,* Claudia Card ed. Lawrence, Kansas: Kansas University Press, 1999

Inviting moral philosophers to re-approach the topics of moral communication and moral community from the perspective of agents who interact frequently or exclusively with others whose moral thinking is prejudiced and parochial, Calhoun's argument calls attention to the relationship between moral possibility and social intelligibility. I will argue that we can appreciate the "deep sociality" of morality without supposing that cooperating in a shared system of meaning is itself a moral ideal that can come apart from the ideal of doing right. I argue further that our communicative exchanges have *afterlives* through which the potential effectiveness of our agency persists. The fact that failures of moral communication may be stages in communicative processes that extend far beyond an initial series of actions and responses supports the conclusion that moral innovators' misread actions are best understood as *incomplete* (yet completely remarkable) moral successes.

Calhoun is careful to clarify that the type of moral failure she identifies is not blameworthy and does not signal any wrongdoing. On her view, the possibility of moral failure produced by acting well stems from the existence of two moral ideals that can and do come apart: the ideal of doing the right thing according to an objective moral standard and the ideal of participating in a shared social scheme of moral meaning. The former ideal is familiar, but Calhoun must provide an argument for the claim that participation in a shared social scheme is itself a moral ideal. She does so by appeal to the social nature of morality as a practice, an argumentative strategy very similar to that used by Margaret Urban Walker in developing her expressive-collaborative model of morality.[116]

Those who resist mistaken moral beliefs that are dominant in their communities should not understand themselves as rejecting social norms in favor of moral norms, since no point of view on the content of morality can be developed apart from social experience. According to Calhoun, the more accurate contrast would be between a hypothetical social morality that an agent endorses and the actual social morality that she rejects from the perspective of the hypothetical, more fully justified, moral order. Because we enter into a moral practice that is already underway, and because the aim of this practice is to "to make our common lives good," we cannot make a distinction between moral standards on the one hand and (mere) social standards on the other. Successfully acting in accordance with the objectively correct moral standard, which is a hypothetical social morality, can lead to failures of intelligibility within the actual social morality in practice.[117]

Although I wish to consider Calhoun's work more as a spur to thought than as a target for objections, here I note that her argument does not clarify the connection between the reason for failed reception and the mark of blameless moral failure. If I interact with a person who, because of idiosyncratic character flaws, does not accept my apology or perceives my gift as an

[116] See Margaret Urban Walker's *Moral Understandings,* Routledge: 1998

[117] The situation that Calhoun initially asks us to consider is that of moral resistors; however, many of her arguments appear to depend on the more narrow case of the isolated or solitary moral resistor. Though there have surely been people in moral isolation, I think that we should distinguish between moral isolation and participation in moral counter-cultures in analyses of moral failure as communicative failure. Agents who resist dominant moral understanding but do so as participants in emerging or established counter-cultures succeed in cooperating in a shared social scheme of meaning and, if they truly have moral insight, do right and act virtuously. We should recognize that there are multiple social moralities in practice, some of which are closer to embodying objective moral standards.

imposition, do I fail to express remorse or manifest generosity in exactly the same way as I would if the bad reception is due to widely shared, but mistaken moral understandings or due simply to cultural differences? I do not think we that we should view communicative failures of these three kinds in the same way and I do not see how the ideal of cooperation in a shared scheme of meaning offers a way to distinguish between them besides, perhaps, frequency of failure. However, it is difficult to see why frequency of communicative failure should give us reason to reframe actions that in isolation would be moral successes, or morally neutral, as moral failures. I suggest that a generally Kantian approach revised to acknowledge relational aspects of autonomy and the conditions of socially embedded rational agency, may be of some help here because it allows us to capture the moral significance of social meanings without introducing a second, competing moral ideal.

A Single Moral Ideal

I appeal to Barbara Herman's interpretation of the Kantian ideal of a kingdom of ends in order to illustrate a theoretical framework in which cooperation in a shared system of moral meaning does not in itself constitute a moral ideal that might come apart from that of doing right. On Herman's reading, the kingdom of ends is not simply an order of beings falling under moral law or a union of good wills, but rather a representation of the moral law as a model of human social order that truly "amplifies the normative content of the categorical imperative."[118] In claiming that the kingdom of ends is best understood as the representation of the moral law as an individual thing, a model social order to be imitated, Herman calls us to think about the implications of conceiving of persons "as in a social union that is, as such, expressive of their rational natures."[119] According to Herman, this social order "is not an order imposed on agents, but an order of agents whose rationality is essentially expressed as much through social as through natural-physical means."[120] Sharing moral understandings with others is constitutive of our ability to act in ways that respect and support our own and others' rational agency. The kingdom of ends guides judgment by directing us to work to develop shared understandings where they do not already exist. This understanding of the kingdom of ends amounts to a single moral ideal that figures sharing moral meanings as a part of the moral vocation of deeply social rational beings. Our interest in doing right is not separate from our interest in acting in ways that reflect an understanding of the social meaning of our actions, including speech-acts and symbolic gestures.

Sharing meaning matters for the same reasons that doing right matters: we aim to share meaning because that sharing is partially constitutive of the expression of our personhood. Hence, the Kantian ideal of doing right is itself a social ideal, the ideal of the kingdom of ends. This perspective on Kantian ethics takes seriously the fact that the practical efficacy of one's moral agency, actually expressing one's volitions in action and realizing one's moral ends, is not completely within one's control and requires more than a good and strong will.

For example, Barbara Herman agrees with Cheshire Calhoun that we cannot separate moral requirement from social practice and she maintains that this fact reveals the mistake in thinking that "whatever morality requires of us as individuals, it will be something that we are, as

[118] Barbara Herman, *Moral Literacy*, Harvard University Press: 2007, p.77
[119] Herman (2007), p.66
[120] Herman (2007), p.72

individuals, able to do, or able to do to the degree that we are virtuous or good."[121] When I travel to a foreign country or teach students who I know come from diverse cultural backgrounds, I know that I will need to make an effort to understand how my actions or the examples I use will resonate. As Herman explains, "in circumstances in which one knows or should know that divergent cultural facts affect the meaning of social gestures, one is under obligation to make one's maxims responsive to those facts."[122]

Herman's claim that concern for moral efficacy should motivate "mutual accommodation" of differences for the sake of extending the community of moral judgment, though very sensible as a response to cultural pluralism, may be puzzling in relation to cases of moral innovation or "resistance". Does a woman a who is seen as nag or ungrateful when she insists on help with housework not fail to make her maxims responsive to social facts in the same way as a traveler in Nepal who fails to do so when she carelessly steps over peoples' outstretched feet in the airport waiting area?[123] I believe that Herman, rightly, does not think so, for judgment in terms of the kingdom of ends does not require us to adjust to morally flawed socio-cultural conventions. Although meaning well does not ensure moral success, we are not vulnerable to moral failure caused by *unreasonable* negative responses to our actions.

> … the moral adequacy of an action can depend on the structure and content of the maxims of other persons. That I act from a maxim of beneficence does not guarantee that I act beneficently. If the recipient of my good will is insulted by what I would do, and *if this response is at all reasonable*, then my action has failed to be the kind of action I willed. This is not a challenge to my moral worth; it calls into question the efficacy of my agency in my action. What makes this of special concern is the possibility that the efficacy of my agency may depend on factors over which I have no complete control and into which I have no automatic insight [emphasis added].[124]

Here Herman proposes that the efficacy of our agency is a function of how completely our actions match the conception of our actions *as we will them*. She holds that when our maxims do not register the *reasonable* social meanings through which others interpret our actions we are vulnerable to moral failure according to the standards articulated by the three formulations of the categorical imperative.

Failures of reception caused by a) a lack of cross-cultural understandings or b) one's own psychological idiosyncrasies, and those caused by c) others' flawed moral understandings or d) others' psychological idiosyncrasies all seem to require different analyses. An action I intend as an expression of sympathy might fail to register as such because a) I do not understand the importance of interpersonal boundaries in the other's culture, b) I am defensive, c) the other considers me to be morally inferior or d) the other suffers from paranoia. The cause of the failure of reception makes a difference to the implications the miscommunication has for one's moral efficacy, understood in terms of fit between action-as-reasonably-interpreted and action-as-willed. The efficacy of one's agency is undermined in cases a) and b), but not in cases c) and d).

---

[121] Herman (2007), p.51
[122] Herman (2007), p.76
[123] Stepping over others' feet is considered a sign of disrespect in areas of Nepal.
[124] Herman (2007), p.36

In response to this approach, we might join Calhoun in asking why the reasonableness of the recipient's response makes a difference to my moral success. Can my action be the kind of action that I willed if it is not intelligible to others, regardless of their reasonableness? Put otherwise, does my action truly *express* my volition if it is not intelligible to those with whom I interact? Further, though I may have willed correctly, surely I have not realized my moral end if I set out to offer aid and instead add insult to injury. Herman's position requires an explanation of why successful expression does not require successful reception, as well as clarification of the relationship between the expression of one's volition in action and realization of one's moral end, both of which are relevant to assessments of moral efficacy in Kantian terms.

A satisfying response to Calhoun's challenge requires further analysis of the nature of moral communication in general, and the impact of probable failures of reception on the formulation of an agent's maxims of action in particular. Both Kant and Kantian theorists have focused on communication from the perspective of the agent who must endeavor to overcome or avoid moral egoism or parochialism.[125] Theorists interested in moral innovation must re-approach the topics of moral communication and moral community from the perspective of agents who interact frequently or exclusively with others whose moral thinking is prejudiced and parochial. While I cannot offer anything like an original theory of either moral communication or moral community here, I will explain some distinctive features of an approach to moral innovation and efficacy based on a revised Kantian ethics.

The Risks of Innovation

Kant understands the moral egoist as one who determines his will solely on the basis of his own happiness, without concern for the objective validity of his maxims: "the moral egoist is a man who limits all ends to himself, sees no use in anything except what is useful to him."[126] He does not consider the possibility of another kind of moral egoist, namely one who, though concerned with action on principle, misinterprets the application criteria for principle and insulates himself from correction by others. To identify this possibility as a problem for the Kantian agent is to deny that reason provides individual agents direct access to the application criteria for moral principles. Some Kantian theorists who articulate this position understand it as a correction to Kant's implausible individualism, while others present it as part of an unorthodox view of Kant's own theory. Richard Eldridge, for example, contrasts his view of principled action to Kant's:

> Crucially, and contrary to Kant, the establishment and testing of both the application of criteria of respect and the value of acting according to them *cannot* be the work of individual consciousness or conscience alone… Criteria are established publicly, as they are articulated and lived out by members of a form of life.[127]

Barbara Herman agrees with Eldridge, though she considers the socially embedded nature of rational agency to be presupposed by Kant's account of judgment in terms of the categorical imperative test procedure. She maintains, "[A]lthough autonomy is an essential property of individual rational wills, for human beings, autonomous moral agency is realized in and through

[125] See for example Onora O'Neill's "The Public Use of Reason" in *Constructions of Reason* (1989), and Allen Wood's "Reason, Communication and Enlightenment" in Kant's Ethical Thought (1999).
[126] *Anthropology* [7:130]
[127] Eldridge (1989), p.36

a certain form of life with others"; indeed, "we acquire our most basic moral concepts… as part of a social practice."[128]

With Eldridge and Herman, I hold that it is possible for individual agents to revise and extend the criteria for expressing respect for persons, training in which constituted the development of their rational autonomy. Eldridge and Herman worry about the dangers of unintelligibility and moral egoism that accompany imaginative agency, dangers that border the territory of moral failure that concerns Calhoun. According to Eldridge, "[I]ndividuals have a central role in creatively envisioning how to extend the requirements of principle to new cases," however, "whether such creative extensions are apt will itself be a matter for others also to judge."[129] Moral innovators run the risk of "prideful misinterpretation" of principle if they ignore the actions and responses of others, so one cannot forgo concern for successful communication. Similarly, Herman warns, "Although through experience and reflection we may extend or modify our moral lexicon, we risk loss of moral intelligibility if we set out too much on our own."[130]

Let us summarize the risks involved in the kind of individual moral imaginativeness posited by Eldridge, Herman and myself. The morally imaginative agent risks moral error in the form of an interpretation of principle that others understand, but which they justifiably reject. The imaginative agent also risks two forms of moral unintelligibility: unintelligibility to oneself, which entails unintelligibility to others; and, unintelligibility to others, or failure of reception, that does *not* entail erroneous interpretation of principle and does not fully undermine one's ability to make moral sense to oneself. Calhoun urges us to recognize the second kind of unintelligibility as a form of moral failure distinct from the first two kinds. The risk of complete failure of reception, moral isolation, is importantly different from wide failure of reception. Agents who resist dominant moral understandings as participants in emerging or established counter-cultures successfully cooperate in a shared scheme of social meaning.

Herman does not distinguish between the two forms of unintelligibility and Eldridge does not reflect on the possibility of being justified, but wholly misread, in one's imaginative departure from established moral understandings. These omissions may have a common source in the view that complete failure of reception inevitably calls one's intelligibility to oneself into question. Elaborating his view of others' responses as tests of the aptness of imaginative moral moves, Eldridge claims, "[A] grasp of principles and their requirements can be attained only cooperatively and conversationally."[131] If this is the case, then complete failure of reception undermines moral innovation. Further, this view suggests, rightly I think, that even wide failure of reception likely loosens the moral innovator's grasp of principle in the sense that she may experience deep doubt and loss of confidence in the justification of her claims.

A version of Eldridge's claim is consistent with the existence of morally imaginative agents who are right, reasonably confident, intelligible to themselves *and* frequently misread by others. Hence, Kantian theorists interested in the efficacy of agency in concrete social

---

[128] Herman, 2007, p.130; p.144
[129] Eldridge, 1989 p.36
[130] Herman, 2007, p.144. Both Eldridge and Herman appreciate points that are typically associated with Hegel. Hegel's views on recognition and his discussions of conscience and the beautiful soul are all highly relevant to my topic, but I cannot explore these connections here.
[131] Eldridge (1989), p.52

circumstances should be concerned with how moral failure, as described by Calhoun, impacts moral efficacy. Because communication and the development of moral community are ongoing processes, I believe that Kantians should not recognize failures of reception as moral failures. Concern for correctness and concern for colloquy coexist uneasily in the case of moral innovators who resist flawed social norms, but not so uneasily as to create the potential for simultaneous success and failure according to a single, though complex, moral ideal grounded in the value of rational agency.

Communication, Community and Control

Moral innovators present their communities with the task of accommodation. They likely make those with whom they interact uncomfortable, their actions may disrupt social routines facilitated by shared meanings that impede mutual acknowledgement of the value of persons. If the widely shared understandings on offer in one's social world obscure the dignity of persons, constrict communities of judgment or are inadequate to newly emergent circumstances, the Kantian moral ideal does not recommend communicating in terms of social norms as they stand. This is not because communication does not matter or because Kantians do not consider the correct exercise of reason to be deeply social. Attention to i) the ways in which communicative processes extend beyond discrete moments of interpersonal interaction ii) the role of expectations of failed reception in the formation of maxims of action and iii) the incomplete nature of moral understanding in general supports the conclusion that moral innovators' meritorious, yet misread actions are *incomplete* (yet completely remarkable) moral successes.

Failures of moral communication, where they are caused by one party's unreasonable prejudices, may be stages in communicative processes that extend far beyond the initial series of actions and responses that might be deemed a "failed moral communication." Stories of parents who shun their gay children and accuse them of ingratitude and selfishness are not uncommon, yet some overcome this stance and ask for their children's forgiveness. The point is this. Our communicative exchanges have *afterlives* through which the potential effectiveness of our agency persists. This fact complicates the question of whether one has successfully expressed, say, self-respect or gratitude in a given interpersonal exchange. It is similarly relevant to judgments about the realization of moral ends in cases of failed reception.

Calhoun's stance on the question of expression has awkward implications that can be resolved by panning out, as it were, to a wider view of moral communication. On her view, the action of a young man who comes out to his family *counts as* selfish if his family so interprets him. Presumably, his choice becomes unselfish as soon as his interpreters' understanding shifts. I think Calhoun commits herself to this position because she wants to capture something important about the limits of individual agency. It may be true that, during his parents' initial period of emotional response, nothing the young man could do in good faith would register as an expression of gratitude.

In the kinds of cases of failed reception that interest Calhoun, the agents' expressive success is *incomplete.* Here I want to distinguish between a partial success and an incomplete success: were I to claim that agents partially succeed in the expression of virtue when reception fails, this would imply that they also experience a partial expressive/moral failure, and my position would not diverge from Calhoun's. The expressive efforts of moral innovators cannot be appropriately judged in isolation from the communicative processes they may initiate or

continue, so there is an open-endedness to the question of their intelligibility. We may complete morally imaginative agents' moral successes when we listen to them and acknowledge the sense in their justifications, when we read and respond with understanding to the stories they write, when we come to interpret their actions as expressive of shared values. Possibilities for moral meaningfulness that have been forgotten can be reanimated and adapted to different contexts by agents who "inherit" expressions that were once ill-received.

Secondly, where an agent expects that her action will be misread by others, this expectation enters into her conception of the action as choiceworthy (and hence into her maxim of action). This does not mean that an expectation of bad reception will necessarily deter the agent from acting as she would without such an expectation. An agent may believe that her perspective on value warrants accommodation by others and hope that her interlocutors will become capable of this. An agent may continue to expose herself to bad reception because she understands herself as a member of an imagined moral community that sustains her sense of intelligibility. An expectation of failed reception may prevent an agent from attempting communication with some others, but it may not deter an agent who conceives of her action as the only way to create the possibility of change.

Finally, the idea that we do not possess complete moral understanding is a deep Kantian trope that reframes all moral success as incomplete success, marked by individual epistemic limitation. In his writings on enlightenment and history, Kant emphasizes that our use of reason is progressive: "Reason itself does not work instinctively, but requires trial, practice, and instruction in order gradually to progress from one level of insight to another."[132] Kant proposes that each generation "passes its own enlightenment to its successor…" and submits that Nature "reveals something, but very little" of a path towards a kingdom of ends.[133] Kant's understanding of morality as the historical vocation of a community that spans generations leaves great room for moral innovation, but also stresses the limits of individual success.[134] All persons gain epistemic access to the criteria for respect for persons through enculturation. Unfortunately, or "perversely" as Kant would have it, we are led "from culture to morality, and not (as reason prescribes) from morality and its law, as the starting point, to a culture designed to conform with morality."[135] We must move *through* culture towards full respect for the dignity of humanity. The incomplete successes of moral innovators help explain how movement in this direction is possible at all.

---

[132] Immanuel Kant. "Idea for a Universal History from a Cosmopolitan Point of View", in *Kant on History*, Lewis White Beck, trans. The Bobbs-Merrill Co., 1963, p13. For discussion of Kant's view of reason as historically situated in its use and yet unchanging in its principles, see Chapter 7, section 1 of Allen Wood's *Kant's Ethical Thought*, Cambridge University Press: 1999. Wood notes that for Kant, "…reason also has the capacity to transcend its situation, to generate higher universal standards (which Kant calls 'ideas') through which social traditions may be criticized and changed" (p. 230).

[133] Ibid.

[134] The principles of pure practical reason are a-priori and so moral innovation does not take place at this level. Rather, there is great room for innovation with respect to conceptualizing the criteria for application of principle in specific socio-cultural and institutional contexts.

[135] Immanuel Kant. *Anthropology from a Pragmatic Point of View* (1798), translated by Mary J. Gregor. The Hague: Martinus Nijhoff, 1974, p. 295, 7:328.

In closing, I would like to emphasize that I agree with Calhoun's claim that moral efficacy is not totally within one's control. Others' failures to understand our justified claims and expressions of moral understanding may limit our moral efficacy and may even force us to abandon certain ends and develop new routes to self-expression.

Alex Worsnip
Yale University

Alex Worsnip is a fourth-year PhD student at Yale University. He works primarily in epistemology and in the theory of normativity (both broadly construed), as well as related areas of philosophy of language and mind, and of first-order moral and political philosophy. He has articles forthcoming in journals such as *Philosophers' Imprint* and *Pacific Philosophical Quarterly*.

"Cryptonormative Judgments"

Commentator: Raff Donelson, Northwestern University

Abstract. *A cryptonormative judgment, roughly speaking, is a judgment which is presented by the agent who makes it as non-normative (either generally or in some particular respect), but which is in fact normative (either generally or in that particular respect). The idea of cryptonormativity is familiar from debates in social theory, social psychology, and continental political philosophy, but it has to my knowledge never been treated in analytic metaethics, moral psychology or epistemology except in passing. In this paper, I argue, first, that cryptonormative judgments are pervasive: familiar cases from everyday life are most naturally diagnosed as cryptonormative judgments. Second, they reveal that normative judgment is a state which can be quite deeply non-transparent to its bearer, in a way that is not, for example, assimilable to the phenomenon of self-deception. Third, they shed light on debates over amoralism and lend some support to a picture of normative psychology that links normative judgment constitutively to motivation.*

Rex, the secret agent. Rex is a secret agent for a major Western nation. He has captured a terrorist suspect whom he believes to know the location of a ticking bomb. Short of time, and unable to extract information from the suspect by conventional means, Rex turns to less conventional methods that border on torture. Rex's partner agent protests, "you shouldn't do this – it's morally wrong!" Rex replies, "I don't give a damn about the morality of it. The most important thing is that we save the millions of innocent lives that are in danger right now."

Roland, the sexist headhunter. Roland is a headhunter who recruits people for major financial firms. Anyone who observes Roland's headhunting practices can tell that he tends to favor hiring men over women, other things being equal. Confronted about this by a client, Roland frankly admits that he prefers not to hire women. His client asks him, "how can you justify such a sexist attitude?" Roland shakes his head knowingly, having had this conversation hundreds of times before, and replies, "you have to understand – I'm not making any value judgment here. It's just how things are – the rough-and-tumble of the financial world is no place for a woman. It's just a biological fact that they're not suited to it. I wish it were different – but we have to be realistic."

Christine, the modernizing politician. Christine is a politician in a Western democratic state. She belongs to a traditionally left-of-center party which used to be quite radical, but went through some rough times with the electorate, and she is part of a small cabal of senior figures within the party who are intent on modernizing it and bringing it toward the center-ground of the nation's politics. One of the things she particularly wants to avoid is what she calls the 'ideological' character of the party's previous leadership. She states that she believes that modern politics has moved on from ideological dispute and arcane philosophical debates about justice and equality, and that the party now has to simply focus on being pragmatic and doing what works.

In my view, the cases of Rex, Roland and Christine all exemplify a single phenomenon which I call *cryptonormative judgment*. Very roughly, a cryptonormative judgment is a judgment which is presented by the agent as non-normative (either generally or in some particular respect), but which is in fact normative (either generally or in that particular respect).

The idea of cryptonormativity is familiar from debates in social theory, social psychology, and continental political philosophy,[136] but it has to my knowledge never been treated in analytic metaethics, moral psychology or epistemology except in passing. This is somewhat surprising, since cryptonormative judgments are familiar and pervasive features of ordinary life. In this paper, I hope to show that cryptonormative judgments are not only philosophically interesting in and of themselves, but that they are shed light on extant debates about the conditions for making normative judgments, as well as for being in mental states more generally.

In part 1, I will give a definition of 'cryptonormative judgment' and makes some remarks about it. In part 2, I will give my basic diagnosis of the cases we began with as cryptonormative judgments, and suggest that such cases are pervasive. In part 3, I consider several possible attempts to reject the diagnosis, and reject those attempts. In part 4, I argue that the phenomenon of cryptonormative judgment shows us that normative judgment is a mental state that can be deeply non-transparent to its bearer. In part 5, I draw some lessons from this conclusion for existing debates in metaethics and moral psychology.

1. Defining 'cryptonormative judgment'

Although cryptonormative judgments are, I have just said, pervasive, 'cryptonormative judgment' is far from a folk term. I am thus using the term stipulatively here. First, let me be clear about how I am using 'judgment'. Sometimes 'judgment' is used in philosophy in a way that is supposed to signal an occurrent act of the mind, as opposed to a settled and permanent state. However, in metaethics, 'judgment' is not generally used in this way. Rather, 'judgment' is used as a kind of placeholder term for a normative mental state that is intended to be neutral as

[136] See, e.g., Habermas (1998); Anderson (1992); Kolodny (1996); Kamolnick (1998); King (2009); Maes *et al* (2012).

so whether such mental states are ordinary beliefs, some special kind of belief, pro-attitudes, some hybrid of the above, or some other mental state.[137] I am using 'judgment' in this latter way.

That clarified, here is my stipulative definition of a cryptonormative judgment:

Cryptonormative judgment. A cryptonormative judgment is a judgment an agent makes such that

(i) The judgment's content is at least partially normative in character
(ii) The agent sincerely denies that she makes the judgment (presented so that its particular normative content is made explicit)
(iii) The agent sincerely describes the judgment such that, if her description were accurate, the judgment would either (a) lack normative content entirely or (b) lack a particular kind of normative content which the judgment in fact has.

A few remarks on this definition.

First, the definition does not require that the agent have the concept *normative judgment*. This would be the case if we defined a cryptonormative judgment in terms of the agent's sincere denial *that she makes a normative judgment*. We don't want this result, since *normative judgment* may be a concept that many ordinary people lack. Fortunately, the simplest cases of cryptonormative judgment clearly do not require the agent to have this concept. Suppose someone judges that one ought to Φ, but denies that she judges that one ought to Φ. When asked to explain her views with respect to Φ-ing, she simply presents her judgment as the belief that Φ-ing is what most people do. Her judgment then meets all the above conditions, but clearly we have not required that she possess the concept *normative judgment*.

That said, the definition will also cover more sophisticated cases where *normative judgment* or a closely related concept (such as the more common *value judgment*) is directly employed by the agent. Suppose an agent affirms that Φ-ing is common sense, but denies that he makes any 'value judgment' thereby. Unlike judging that something is what most people do, there is a case to be made that judging something to be common sense *is* a normative judgment.[138] But even though the agent says that Φ-ing is common sense, the fact that he explicitly says that his judgment is not a value judgment means that if his description were accurate, the judgment would lack normative content. So the case meets condition (iii). And once the normative content implicit in the claim that Φ-ing is common sense is made explicit, he would deny that he makes any such judgment, so he also meets condition (ii). So this kind of case is also covered by the definition.

---

[137] See, e.g., Sturgeon (2007).

[138] A different interpretation of the case has it that *Φ-ing is common sense* is not itself a normative judgment, but that under many conditions those who rely on the claim that Φ-ing is common sense are usually also making a further normative judgment that Φ-ing ought to be done. In that case, this example meets the definition of cryptonormative judgment in the same way that original simple case did – though with respect to the judgment that Φ-ing is to be done, not with respect to the judgment that Φ-ing is common sense. Nevertheless, the more sophisticated structure seems possible in at least some cases.

Clearly the simpler and more sophisticated case have some differences, but they both exhibit the same phenomenon of interest for our purposes. So I have no compunctions about covering them both under the same definition.

Notice also that the definition allows for a judgment to be cryptonormative if it is sincerely presented by the agent as having some kind of normative content, but a different kind than it actually has. So, for example, an agent might present a moral judgment as being a judgment of prudence, and this could count as a cryptonormative judgment. It might be proposed that such a case is better described as a 'cryptomoral judgment', with 'cryptonormative' reserved for cases where the judgment is presented as entirely non-normative. But I think that the cryptomoral case is interesting in just the same way as the more narrowly cryptonormative case, and so once again I am happy to include both under the definition of 'cryptonormative', which is in any case a term of art.

Finally, notice the sincerity condition in (ii) and (iii). Some social theory literature on cryptonormativity may use a definition which effectively omits this sincerity condition. So, views, stances or theories may be described as 'cryptonormative' for describing themselves in non-normative terms even when this is an intentional act of deception. This way of operating is understandable from a political point of view, but such cases are less interesting from the point of view of thinking about mental states, since the possibility of deliberately misreporting or lying about your mental states is not in any way a puzzling one. That is not to say that the narrower notion which is of interest to us is not also politically interesting, however. In fact, if anything, misleading speech acts which are not based on deliberate attempts to mislead can be all the more insidious in their political effects.

2. The ubiquity of cryptonormative judgments

Here is what I think is the natural diagnosis of what is happening in the cases of people like Rex, Roland, and Christine.

Someone like Rex, the secret agent, thinks of himself as unconcerned with the niceties of morality. Perhaps this is a self-image he has built up over many years. Moreover, his idea of someone concerned with morality is one of someone who is weak, like his partner agent – unable to do what must be done because of emotional attachment. It may be of someone who is out of touch with the realities of the world, or who thinks too theoretically or abstractly. Perhaps Rex even implicitly thinks that to the extent one is concerned with morality, one is committed to a kind of absolutism that is indifferent to consequences or outcomes. But what Rex does not realize is that his own judgment that saving the millions of innocent lives at risk is more important than refraining from torture is in fact itself a moral one. The judgment effectively concerns the relative moral priority of two considerations. So Rex misdescribes his own mental state. It is a moral judgment, even though he sincerely thinks that it isn't. As such, it fits our definition of a cryptonormative judgment.

Roland, the headhunter, goes further that Rex: he denies making any kind of value judgment whatsoever. According to Roland, all he has is a descriptive belief that women are not suited to the rough-and-tumble of the financial world: no kind of normative judgment comes into

it. But again, Roland is wrong. First, the notion of being 'suited' in this context has normative content. There is no straightforward set of criteria for what it is to be 'suited' to a workplace that can be given without saying what a *suitable* worker in that workplace would be like. Second, even if we were to concede to Roland that 'women are not suited to the financial world' is a purely descriptive judgment, Roland would still be making a normative judgment in taking this purported descriptive fact to have some kind of upshot for who should be hired. Roland, too misdescribes his own mental state. He makes a cryptonormative judgment.

Christine, the politician, is similar to Roland, but in an explicitly political context. She assumes that there is a value-neutral characterization of 'doing what works' that avoids the need for any kind of philosophical or ideological framework. But what she describes as 'working' will smuggle in her own value judgments about what the goals of policy should be – what counts as a policy 'working' – which likely reflect particular normative assumptions that she takes for granted, without realizing that she is making them. Again, she makes a cryptonormative judgment.

I have picked these three characters to be recognizable, common archetypes. Most moral philosophers have probably found themselves at some point in conversation with someone like Rex, asking them what they do for a living. Such people are fond of making it clear how 'impractical' they think moral thinking is, and profess not to be guided by such judgments. But of course, such people do make many moral judgments. They make moral judgments about the fairness of being required to work overtime, about their duties toward their children and partners, about their entitlement not be overtaxed, and so on in countless other ways. Often they simply do not reflect on these myriad judgments; other times they have explicit rationalizations for how such judgments aren't really moral in nature. Either way, these are all cryptonormative judgments.

Likewise, Roland and Christine should be recognizable. In many professions and in politics, value-disagreements are often stifled by appeals to 'facing facts' or 'facing reality' that encode cryptonormative assumptions about what the normative import of those facts or reality is. Newspaper op-ed columns frequently oppose 'moral' or 'ethical' considerations to the promotion of practically beneficial consequences without explanation.[139] Many people think of rationality or what 'makes sense' as non-normative concepts. And politicians frequently use the notion of 'what works' (and related notions) in precisely the way that Christine does in our example.[140]

Moreover, philosophical debates themselves are not immune from cryptonormativity. The recent 'realist turn' in political philosophy often appeals to a need to eschew 'moralizing' political thought for a kind of 'pragmatism' that is, in my view, cryptonormative.[141] And for many years (though to a far lesser extent now), some philosophers underappreciated the

[139] For just one extremely striking recent example, see Moore (2013).
[140] See, e.g., Blair (2005).
[141] See, e.g., Geuss (2008); Bourke (2009); Williams (2005: ch. 1); Philp (2010). I develop the charge of cryptonormativity against political realism (both in philosophy and in mainstream political discourse) in co-authored work in preparation (Leader Maynard & Worsnip ms.).

normative character of various non-moral judgments such as those concerning rationality, prudence, and epistemic justification – leading to cryptonormative claims in epistemology and ethics.[142]

3. Arguing against alternative diagnoses

But some may wonder if there are alternative diagnoses available for many of these cases. To argue that these cases are not cryptonormative judgments, one will need to hold that they ultimately either (i) don't really count as normative judgments at all or (ii) are not sincerely disavowed by the agent. I will consider several strategies for each route, building from the least challenging up to the most challenging, as I see it. (This takes me in an order that is formally odd, but materially better for the exposition.)

Strategy 1 for resisting (ii): the agent is insincere

One option would be to deny (ii) by denying that the agents in our examples are sincere in their disavowal that they make a normative judgment.

In at least many of these cases, however, is strategy is just too *ad hoc*. Consider Rex, for example. I do not see any reason to think that Rex is deliberately lying when he denies that he makes moral judgments. What reason would he have to do so? It's not like his case against his partner agent fundamentally depends on being presented in non-moral terms. Rex's confusion about what makes something a moral consideration leads to mistakenly thinking that he does not make moral judgments. It's part of his being genuinely mistaken here that he is not being insincere. Now, there may be some politicians who are deliberately disingenuous in presenting their agenda as value-neutral. But it seems to be an unwarranted empirical assumption to assume that every politician who presents their agenda this way is being insincere.

Strategy 1 for resisting (i): the agent is an amoralist (or anormativist)

Remember that we said that Rex thinks that moral considerations are absolute in a way that makes them blind to consequences. It might be thought that it just follows straightaway from this that Rex does *not* make the judgment that torture is morally justified. For Rex, one might think, has a particular view of morality, one on which morality consists of absolute prohibitions and is blind to consequences. Even if this moral view is mistaken, the argument goes, it might still by Rex's view. It's just that, as he says himself, Rex does not care about morality (perhaps partly *because* of his absolutist moral views!) He is an amoralist. We can therefore say that he is not mistaken about his own mental states.

This is not a plausible diagnosis of Rex's state of mind. First, if Rex's judgment that the lives of millions of innocent people are more important than the rights of the suspect to be

[142] For a striking example, see Mackie (1977: ch. 8), who having declared that all value-claims are false, goes on to consider which (fictional) system of morality would be most 'practical', in a way that is supposed not to itself rest on any such value-claims.

protected from torture is not a moral judgment, then what is it? It is certainly not a judgment about what is prudentially good for Rex. It seems to play the exact functional role of a moral judgment – justifying the prioritization of one good over another, and effectively holding Rex's own action accountable to the people whose lives are at stake. Moreover, we can easily imagine Rex being disposed to have the central reactive attitudes associated with a moral judgment that he ought to save the millions: being disposed to feel guilt if he failed to do so; being disposed to blame others (such as his partner) who do otherwise, and so on, without changing the fact that he sincerely denies making a moral judgment.[143] By contrast, he may be disposed not to have any of these functional or reactive associations with the converse judgment.

So Rex simply misdescribes his own state of mind. It's true that he officially avows the view that morality is blind to consequences. But this is not for him a substantive view which he has about what morality requires; rather, he thinks (wrongly) that it's something like constitutive of something being a moral consideration or judgment that it not concern itself with consequences. We can see this because he would not classify *others* as making moral judgments that do concern themselves with consequences, and then classify himself as disagreeing with those others. Rather, he would classify any judgment that concerns itself with consequences as a non-moral one.

Rex's actual substantive moral judgment is that one ought to torture the suspect. But because he is confused about what a moral judgment is, he misdescribes his own mental states, and denies that he makes any moral judgment.

Moreover, even if the amoralism strategy worked, it does not generalize in any obvious way to cases like that of Roland, who denies that he makes any normative judgment at all. One might try to say of Roland that he is a kind of general 'anormativist', who ignores normative judgments entirely and simply acts from his desires. But this reply once again passes over the fact that Roland may have all the functional dispositions and reactive attitudes associated with normative judgment. What seems much the better diagnosis is to say that Roland simply doesn't realize that he is making a normative judgment when he characterizes women as ill-suited to the workplace.

Strategy 2 for resisting (ii): linguistic error precludes genuine disavowal

In the light of the previous reply, one may now worry that the cryptonormativity account does not make the agents in question very deeply mistaken about their own mental states after all. For it might be starting to look like Rex (who I'll again use as my main example) just misuses the term 'moral judgment'. Suppose, by analogy, I've been misinformed and think that 'ennui' refers to a state of intense excitement. I might now sincerely assent to the sentence 'I am in a state of ennui', even though the sentence actually expresses a falsehood. But this is a rather uninteresting phenomenon, and it is not a very deep way of my mental states being non-transparent to me! For

[143] For a detailed work on the connection of morality and moral judgment with the reactive attitudes, see Darwall (2006). The inspiration for this kind of approach is from Strawson (1962).

all that has been said, my mental states are fully transparent to me! What I am mistaken about is the meaning of a word, not my own mental states.

In fact, we might even say that in the relevant sense, then, we might say that I don't really claim that I am in a state of ennui. I utter the worlds, 'I am in a state of ennui', but my mistake about the meaning of the word 'ennui' means that I don't really think I'm in a state of ennui at all. And one might worry that the same is true of Rex: his error is merely a linguistic one about the meaning of 'moral judgment'.

However, I think that this too misdiagnoses what is going on with Rex. My claim has not been that Rex doesn't understand the meaning of 'moral judgment'. Rather, it is that Rex doesn't understand what a moral judgment *is*. To try to convince you that these two possibilities are distinct, here is an analogy. Suppose that I announce that I think our friend Tom is depressed, and you announce that you disagree. Suppose also that I am wrong and you are right. There are at least three possibilities here:

(1) You and I understand equally well what depression is, and I am just mistaken about whether Tom is in this state
(2) I lack a full understanding of what it is to be depressed, and thus mistakenly diagnose Tom as depressed
(3) I fail to know the meaning of the word 'depressed', and thus announce something which I do not really believe

What I am insisting on here is the distinctness of (2) from both (1) and (3). (2) is a way of being genuinely mistaken about whether Tom is depressed, even though this mistake is traceable to a misunderstanding of depression in a way that the mistake in (1) is not. (2) is the sort of misunderstanding that might be a result of not really even having been seriously depressed oneself (amongst other causes); unlike the error in (3), it might not necessarily be put right by consulting a dictionary. We could even say that this misunderstanding results in a kind of impoverishment of my concept of depression. But that does not make it a trivial error, in the way that (3) is a trivial error. We would not say in (2) that you and I are just using 'depressed' in two different ways, and that both are right in our own idiolect. (I can have an impoverished concept of something relative to yours, without our talking past, or failing to contradict, each other.) I have made a mistake deeper than that. Moreover, it is *correct* to attribute the belief that Tom is depressed to me; that would not be the case in (3).

We can now cross-apply this to the case of Rex. If Rex really fully understood what it is for something to be a moral judgment, then he would probably classify himself as making a moral judgment. But, as the above analogy shows, that he lacks this full understanding does not just make it the case that he makes a mistake that is trivial or merely linguistic in classifying himself as not making a moral judgment. That ignores the possibility analogous to (2) – that Rex is mistaken about what a moral judgment is (as opposed to the meaning of 'moral judgment'). Just as it was correct to attribute to me the belief that Tom is depressed, it will then be correct to attribute to Rex the belief that he does not make any moral judgment (which would not be correct if he was merely misusing a phrase). Still, it may be that Rex does in fact make a moral

judgment, given what a moral judgment really is. So the possibility of genuine error is allowed for here.

Again, things are even clearer with Roland. Roland might be a perfectly competent user of the term 'normative judgment'. He simply doesn't realize that the judgment that women are ill-suited to the workplace is a normative judgment, because he hasn't thought hard enough about exactly what makes something a normative judgment.[144]

Strategy (ii) for resisting (1): the mental state is alief

In important recent work, Tamar Gendler (2008a, 2008b) has argued that there are mental states that typically move us to action in the way that beliefs do, but which lack the cognitive characteristics of belief. Gendler calls these states 'aliefs'. One cognitive feature of aliefs that helps to explain why they are not beliefs is that we will not typically endorse them as beliefs.[145] So one might think that what I have been classifying as cryptonormative judgments are in fact aliefs. This might encourage a critic to claim that the agents in our examples do not really make the normative judgments I attribute to them: rather, they alieve them. This might be particularly encouraged because Gendler has used the alief framework to analyze implicit bias (Gendler 2011), which reminds us of Roland.

Gendler is surely right that one can be acting as if one believed something without actually believing it. And I am open to the possibility that the notion of alief, or something like it, is what is needed to make sense of this phenomenon. However, I do not think that the cases of cryptonormative judgment that I have described are cases of alief. There is a big difference between Roland's brand of implicit sexism and the sort of implicit sexism that Gendler takes as paradigmatic. Gendler is interested in people who reflectively endorse egalitarian, liberal or feminist sentiments, but end up manifesting their bias unconsciously in action and decision. But Roland is not like this. Roland's bias is essentially overt, insofar as he admits that he prefers to hire women. What Roland denies is that this constitutes any kind of normative judgment. So he is not explicit about his attitude *qua normative judgment*; still, he is explicit about the attitude under at least some description, and he reflectively endorses it: he says that women are ill-suited to the workplace.

Consequently, Roland is not in some way acting contrary to his judgments. Nor does he act in some arational or habitual manner. So he is not an aliever. He is a believer in denial. Gendler's theory points us to the ways in which certain cognitive characteristics are important

[144] At the outset I said that one could make a cryptonormative judgment without having the concept *normative judgment*. But it may look like what I've just said traces a cryptonormative judgment to a misunderstanding about what makes something a normative judgment, and that may seem to require the agent to have the concept *normative judgment*. However, the mistake could simply manifest itself, not specifically in explicit denials that one makes a normative judgment, but in refusal to assent to the content of the judgment presented in explicitly normative terms. In that case one shows an implicit misunderstanding through failure to apply normative terms correctly. (Also, note that in many cases it's not a misunderstanding of what makes something a normative judgment, but a misunderstanding of what makes something a more specific kind of judgment such as a moral judgment. And probably more people have the concept *moral judgment* than have the concept *normative judgment*.)

[145] Though, for a different characterization, on which they are 'in-between' beliefs, see Schwitzgebel (2010).

for bona fide belief, but this does not go so far as to make it impossible to be mistaken about what one believes, and nothing Gendler says suggests that she intends her theory that way.

The same thing holds for Rex and for Christine. For example, Rex reflectively accepts that it is more important to save the millions – he just denies that this is a moral judgment. It's characteristic of a cryptonormative judgment that it is endorsed under some misleading description that is non-normative. This is a fundamental difference with an alief, which is typically not endorsed under any description. Relatedly, cryptonormative judgments are sensitive to factors which their bearers take to be truth-indicative, in a way that aliefs are not.

4. Cryptonormative judgments as failures of self-knowledge

I have argued that cryptonormative judgments are pervasive. Since agents who make cryptonormative judgments are mistaken about whether they make normative judgments, this suggests that normative judgments are mental states which are non-transparent to their bearers.

On its own this may not come as much of a surprise. The strong doctrine that one enjoys genuine infallibility with respect to one's mental states was once popular, counting amongst its adherents Descartes (M III: 37), Hume (T 1.4.2.7) and Wittgenstein (PI: 246-7). Few contemporary philosophers would defend this very strong claim.[146] However, many philosophers nevertheless hold that we enjoy some kind of qualitatively unique privileged access with respect to our own mental states.[147] I venture that many philosophers would side with Donald Davidson (1987: 441) when he writes that there is "an overriding presumption that a person knows what he or she believes." The pervasiveness of cryptonormative judgment suggests that this 'overriding presumption' may be a mistake, at least with respect to the normative domain.[148]

Moreover, I think that cryptonormative judgment illustrates the way in which one can be mistaken about one's own mental states in a quite deep way. In particular, I want to argue that they show that one can be mistaken about one's own mental states without self-deception, which some people might think of as the paradigmatic way to be mistaken about one's own mental states. To do this, I need to argue that cryptonormative judgments are not (necessarily) instances of self-deception. So they cannot be assimilated to self-deception. The attempt to perform such an assimilation would not be an attempt to deny that there are cryptonormative judgments, or that they are not common, as the strategies considered in the previous section were. Rather, the idea is to query how deep mistakes about one's own mental states can go.

In order for the assimilation to achieve this, one needs a notion of self-deception that can be characterized as only involving failure of self-knowledge in some limited or shallow sense. This puts some constraints of the notion of self-deception in play. On one hand, it won't do to just define self-deception in terms of being mistaken about one's own mental states.[149] If it is

[146] An exception may be Burge (1988).
[147] See, e.g., Heil (1988); Davidson (1984; 1987); Shoemaker (2009).
[148] In different ways, numerous philosophers have held that it is a mistake more generally. See, e.g., Ryle (1949: ch. 6); Williamson (2000: ch. 4); Schwitzgebel (2011). For a reply to Williamson see Berker (2008).
[149] C.f. Shoemaker (2009). As well as being too broad, this definition is also too narrow, since it limits cases of self-deception to beliefs about one's own mental states. But clearly one can be self-deceived about matters that do not concern one's own mental states. For example, one can be self-deceived about whether one's long-lost relative is

defined this way, then trivially cryptonormative judgments (and all other instances of being mistaken about one's own mental states) would be instances of self-deception; but this would obviously not show that they do not involve especially deep failures of self-knowledge.

On the other hand, we also won't want a notion of self-deception that precludes cases of mistaken beliefs about one's own mental states from counting as self-deception. For example, Tamar Gendler (2007) characterizes self-deception as a kind of pretense: when you are self-deceived about P, you do not really believe P, but rather pretend (in a particular way) that P is true (in fact, Gendler urges, you may really believe that not-P). But then, if P is some proposition about your mental states, it follows that when you are self-deceived about your own mental states in Gendler's sense, you do not actually have a false belief about your own mental states. So this notion of self-deception is also no good for the purposes of someone who wants to claim that the only way to be mistaken about your own mental states is through self-deception.[150]

What is needed for the assimilator to steer between these two dangers is some notion of self-deception which cashes self-deception out in terms of one's believing P, perhaps due to some motivational bias, despite at some level knowing or being in a position to know that P is false. This vague characterization leaves a lot of room open,[151] but it is enough, I think, to argue that cryptonormative judgments need not involve any self-deception.

First, when you make a cryptonormative judgment, you may not be a position to know that you are making the judgment that you make. Consider again Rex. Rex is precluded from knowing that he makes a moral judgment by his failure to fully understand what a moral judgment is. Consequently, he is not in fact in a position to know that he is making a moral judgment. Of course, one could interpret "position to know" loosely enough that one is in a position to know p when one could, by careful reflection that one may not actually be fully capable of, come to know p. This would then count Rex as being in a position to know that he makes a moral judgment. But this is obviously too loose a gloss on "position to know" to be of use in an account of self-deception. For example, it would count one as self-deceived about complex *a priori* mathematical truths that are beyond one's understanding.

Second, relatedly, cases of cryptonormative judgment do not require the belief that you do not make the normative judgment in question to be due to motivational bias. It can arise from entirely blameless misunderstanding. In some cases it may be that the agent at some level willfully misunderstands, but nothing about the phenomenon of cryptonormative judgment requires this.

It also won't work to try to assimilate cryptonormative judgments to the phenomenon of fragmented belief (a phenomenon closely related to that of self-deception). In recent work,

---

still alive. The over-narrowness and over-broadness here both result, I think, from the same mistake: that of confusing the notion of being deceived *about* oneself with the notion of being deceived *by* oneself.

[150] Perhaps there are good reasons to favor Gendler's characterization of self-deception as pretense, or at least to think that many of the cases that are usually referred to as self-deception involve pretense rather than belief. If so, that just reinforces the problem for executing the strategy of assimilating cryptonormative judgment to self-deception. For cryptonormative judgments are genuine instances of mistakes about one's own mental states.

[151] One characterization of self-deception that fits this mould is that of Mele (1997), but it leaves substantial room for disagreeing with Mele.

Daniel Greco (forthcoming) has appealed to fragmentation to defend the 'iteration principle' that if you believe something, then you believe that you believe it. The fragmentation strategy, applied to cryptonormative judgment, would say that those who make cryptonormative judgments are fragmented in their mental states. Take Rex, for example. On the fragmentation view, in one sense, or relative to one purpose (perhaps the purpose of making declarations), Rex doesn't believe that it is morally justified to torture the suspect. But in another sense, or relative to another purpose (perhaps the purpose of making decisions), Rex does believe that it is morally justified to torture the suspect. The idea is then that, as long as we confine ourselves to one 'fragment' of Rex's mental states, holding the relevant purpose constant, we can then apply the iteration principle. So, given the purposes relative to which it's true to say that Rex believes it's morally justified to torture the suspect, it's also true to say that Rex believes that he believes it's morally justified to torture the suspect. So in an important way, the extent to which Rex can be mistaken about his own judgments here is limited.[152]

However, I do not think that even this more limited supposition is defensible. Suppose we implement Greco's strategy by saying that relative to the purposes of making declarations, Rex doesn't believe that torture is morally justified; but relative to the purposes of making decisions, Rex does believe that torture is morally justified. I do not see what justifies the claim that, even just relative to the purposes of making decisions, Rex *believes that he believes* that torture is morally justified. Attributing this belief is not required to make sense of Rex's decisions or dispositional profile: all that is required is the first-order belief that torture is morally justified. That fact is that Rex has no dispositions of *any* sort that indicate that he *believes that he believes* that torture is morally justified. So there does not seem to be *any* purpose relative to which there is any reason to attribute this belief to him.

Finally, one might wonder whether cryptonormative judgments really constitute failures of *self*-knowledge. In the previous part, I traced cryptonormative judgments in part to failures of understanding of what a normative judgment *is*. But it might be thought that this is not a failure

[152] Greco himself actually thinks that his strategy preserves the thought that our mental lives are sometimes opaque to us, on the grounds that we'll sometimes lack accurate *explicit* beliefs about what our implicit beliefs are, where explicit and implicit belief are fragmented (Greco forthcoming: 13). But it seems clear that this still severely limits how opaque our mental lives can be to us. After all, it precludes inaccurate explicit beliefs about our explicit beliefs, as well as precluding inaccurate implicit beliefs about our implicit beliefs. And ultimately, it guarantees that every time I have a belief, I also have a (true) belief that I have that belief – after all, that is just his iteration principle. On his view, it's impossible to lack that belief. This surely is an important way in which our mental lives are not opaque to us on his view.

The point is even clearer when we move away from a dichotomy between implicit and explicit belief. Greco uses this as a working model for fragmentation to make his points, but has indicated (p.c.) that he thinks it is overly simplified. Rather, he thinks our mental states fragment into as many different compartments as we can have distinct purposes. But if that's the case, then we don't have a single notion like *explicit belief* which can plausibly be the thing that we lack when our mental lives are opaque to us. Rather, it'll just be the case that when I believe something relative to purpose A, I won't necessarily believe relative to purpose B that I believe it relative to purpose A. But why would we think that the right way to see whether my believing something relative to purpose A is transparent to me is to ask whether I believe relative to purpose B that I believe it relative to purpose A? This doesn't seem like a very fair test! Nor does it seem to reveal any significant opacity.

of knowledge of one's own mental states. Rather, it is failure of knowledge of how to describe or classify one's own mental states.[153]

I do not think this distinction holds up. Knowledge of how to describe and classify a subject matter is an important part of comprehensive knowledge of that subject matter. Suppose that an expert biologist and I are both shown a much-enlarged image of a tuberculosis bacterium. Asking 'what is that thing on the screen?', we give different responses. 'It is a tuberculosis bacterium,' replies the biologist. 'It is a long thin purple thing on a blue background', I reply. I take it that it would be absurd here for an observer to react in the following way: 'the biologist and you have equally good knowledge of the thing you are seeing here. It's merely that the biologist knows how to describe it and classify it better than you.' For my failure to describe or classify the image on the screen *constitutes* a failure to understand what it is that I'm seeing. And this failure is not somehow trivial or non-substantive; it is a paradigmatic instance of my lacking substantive knowledge than the expert biologist has.

In that case, the subject-matter of my failure of understanding is the thing on the screen, the image of a bacterium. In the case of cryptonormative judgment, the subject-matter of my failure of understanding is my own mental state. Just as the traceability of the former to my inability to describe and classify images of bacteria does not preclude it from being a failure to know about the image of the bacterium, so the traceability of the latter to my inability to describe and classify my mental states does not preclude it from being a failure to know about my own mental states.

5. Broader metaethical lessons

Given that cryptonormative judgment is a pervasive phenomenon, we can draw an important general lesson: sincere assent to a normative judgment is not a necessary condition for making such a normative judgment. This conclusion has significant ramifications for some existing metaethical debates – in particular, debates about the relationship of normative judgment to motivation, including debates between so-called motivational 'internalists' and 'externalists'.[154]

According to a widespread kind of motivational externalism, whether an agent makes a normative judgment is in no way constituted (not even partially) by whether that agent is motivated to act on the judgment in question.[155] On this view, the agent's being motivated could

[153] Such an objection might be inspired by Burge (1988: esp. 661-63). Burge makes his point in the context of trying to reconcile a kind of infallible self-knowledge with externalism about mental content. He plausibly contends that knowledge that one is in a mental state does not require that one grasp the conditions for knowing that one is in that mental state. I do not have to deny this; what I do deny (and what does appear to put me at odds with Burge) is that knowledge that one is in a mental state might be blocked by failing to grasp the conditions for *being* in that mental state.

[154] 'Internalism' and 'externalism' here are being used to mean something entirely different from what they often refer to in epistemology and the philosophy of mind, where internalism is often a claim about access to one's own mental states and epistemic states. This paper has been putting pressure on such claims of privileged access; the 'motivational internalism' I am going on to offer some support for here, by contrast, concerns the relationship between normative judgment and motivation, and is not only distinct from but essentially unrelated to externalism about mental or epistemic content.

[155] For examples of this view, see, e.g., Brink (1986), Svavarsdóttir (1999) and Copp (2007: ch. 8).

at most be construed as *evidentially* bearing on whether the agent makes the normative judgment in question.

What could be used to motivate this kind of externalist view? One classic way has been to appeal to cases of *amoralist* agents.[156] Amoralists are supposed to be agents who make moral judgments, yet are not at all motivated (even in the slightest) to act on them, since they don't care about morality. Yet of course this description of the amoralist begs the question in favor of the externalist, since the internalist will insist that all such cases must ultimately be ones in which the agent actually is motivated to some extent, or ones in which the agent does not count as genuinely making a moral judgment. In order to persuade some heretofore neutral that there can be amoralists, then, the externalist needs to find a non-question-begging way of describing such an agent – one that latches onto agreed characteristics that such an agent might have without being at all motivated – such that the externalist diagnosis, according to which the agent nevertheless counts as making a moral judgment, seems plausible, and internalist diagnosis, according to which she doesn't count as making a moral judgment, seems implausible and ad-hoc.

This is exactly what externalists try to do. So what agreed, non-question begging characteristics can the externalist use to describe such an agent? Typically, one such characteristic – perhaps the central such characteristic – that externalists appeal to is that of an agent who *sincerely assents* to a moral claim. So, stories of purported amoralists involve an agent who sincerely assents to a moral claim. Externalists claim, plausibly, that we can easily imagine someone who sincerely assents to moral claims without being motivated. And then they use this to argue that given this sincere assent, the internalist diagnosis that no moral judgment is in fact made would be *ad hoc*.[157]

However, if what I have argued here is right, the pervasive phenomenon of cryptonormative judgment shows that normative judgment is a state that can be far from transparent to its bearer. As such, we should not take assent – even sincere assent – to be a good guide to whether an agent makes a normative judgment. True, what I have argued here suggests that sincere assent is not a *necessary* condition for normative judgment – whereas what the externalist needs is for sincere assent to be a *sufficient* condition for normative judgment. But if one can mistakenly think that one does not make a normative judgment – and thus falsely but sincerely deny that one makes a normative judgment – then it seems very plausible that one can also mistakenly think that one *does* make a normative judgment – and thus falsely but sincerely deny that one makes a normative judgment. Just as the natural name for the first phenomenon is *cryptonormative judgment*, the natural name for the second phenomenon is *pseudonormative judgment*. Effectively, what the internalist claims is that supposed amoralists in fact make pseudonormative judgments.[158]

---

[156] All those authors mentioned in the previous footnote appeal to such cases.

[157] I offer a similar diagnosis of the debate between internalists and externalists, and a different way of resisting the externalist's main strategy, in Phillips & Worsnip (ms.).

[158] What R.M. Hare, an arch-internalist, calls an 'inverted commas moral judgment' may be a particular instance of a pseudonormative judgment. See Hare (1952: ch. 11).

Externalists typically claim that it is *ad hoc* to say that so-called amoralists do not really make moral judgments. But if they are willing to countenance the phenomenon of cryptonormativity – as I have argued we all should – then they should also be willing to countenance the phenomenon of pseudonormativity without its seeming bizarre, or an *ad hoc* notion to invoke. In fact, the explanations of the two possibilities run in a strongly analogous way. In the case of cryptonormativity, an agent's understanding of what normative judgment is is overly narrow, thus leading to a misclassification of the agent's mental states. In the case of pseudonormativity, the agent's understanding of what normative judgment is overly broad. So, for example, on the internalist story, the putatively amoralist agent might err in self-classification because he thinks that to judge that some act is morally wrong is merely to judge that it falls into the category of behavior commonly described as wrong by ordinary people. But if the internalist is right, this does not suffice: really judging that some act is morally wrong regards seeing it in such a way that one is at least somewhat motivated to refrain from performing it. No wonder, for the internalist, that sincere assent can come apart from normative judgment, since the agent's sincere assent rests on what, by internalist lights, constitutes a misunderstanding of what normative (or moral) judgment is.

Indeed, in many purported amoralist cases, I think it is quite natural to claim that the agent may make both a cryptonormative *and* a pseudonormative judgment, and the attribution of the former strengthens the case for the latter, thus lending support to internalism. Go back to Rex, for example. As well as sincerely denying that he make the moral judgment that torturing the suspect is right, Rex might well also sincerely assent to the proposition that torturing the suspect is wrong. Earlier I argued that Rex's real moral judgment here is that torturing the suspect is the right thing to do: that is a cryptonormative judgment. We could say that Rex contradicts himself by also judging that torturing the suspect is the right thing to do. But in general it seems that the same considerations that favor counting him as judging that torture is right also favor not counting him as judging that torture is wrong. He thinks he makes the latter judgment, but in fact makes the former. So he makes both a pseudonormative and a cryptonormative judgment.[159]

The phenomenon of cryptonormative judgment, then, weakens the externalist and strengthens the internalist in the dialectic familiar from the literature about normative judgment and motivation. It does not, of course, constitute a knock-down proof of internalism. But I do think that the phenomenon of cryptonormative judgment lends the internalist position some positive support. If it turns out that sincere assent is less important for normative judgment than we thought, we are left wondering what else, if anything, determines whether an agent counts constitutively as making a normative judgment. And the obvious candidate for this role is whether the agent is motivated to comply with the judgment in question. After all, part of what makes it plausible to classify the agents we began with as making normative judgments despite

[159] Strictly speaking, the pseudonormative judgment that Rex makes is not that torturing the suspect is morally wrong, since he does not make that judgment at all. Rather, it is the actually non-normative judgment that Rex actually makes but presents as the normative judgment that torturing the suspect is morally wrong. This might be, for example, that torturing the suspect is the sort of thing that most moralistic people call 'wrong'.

their denials that they do is their dispositional and motivational profile. To crudely oversimplify, with these agents, it is not what they say but what they do that matters in attributing normative judgments to them. And clearly this fits in nicely with the internalist picture.

Conclusion

I have argued for three claims in this paper. First, cryptonormative judgments are pervasive: they are the natural way to diagnose familiar cases from everyday life. Second, they reveal that normative judgment is a state which can be quite deeply non-transparent to its bearer, in a way that is not, for example, assimilable to the phenomenon of self-deception. Third, they shed light on debates over amoralism and lend some support to a picture of normative psychology that links normative judgment constitutively to motivation.

No doubt there is much more to be said about cryptonormative judgments. However, I hope to have shown that they are a philosophically interesting phenomenon, worthy of attention from those interested in normative judgment and in mental states more broadly. Even those who will disagree with some of my claims about cryptonormative judgments need to account for this pervasive human phenomenon.

References

Anderson, A. (1992). 'Cryptonormativism and Double Gestures: The Politics of Post-Structuralism,' *Cultural Critique*, 21: 63-95.

Berker, S. (2008). 'Luminosity Regained,' *Philosophers' Imprint*, 8/2.

Blair, T. (2005). Speech to the Labour Party's 2005 conference in Brighton. Published in full at BBC News online, dated 27 September 2005.

Bourke, R. (2009). 'Theory and practice: the revolution in political judgement,' in Geuss & Bourke (eds.), *Political Judgement: Essays for John Dunn.* Cambridge, UK: Cambridge University Press.

Brink, D.O. (1986). 'Externalist Moral Realism,' *Southern Journal of Philosophy*, 24: 23-41.

Burge, T. (1988). 'Individualism and Self-Knowledge,' *Journal of Philosophy*, 85/11: 649-63.

Copp, D. (2007). Morality in a Natural World. Cambridge, UK: Cambridge University Press.

Darwall, S. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Davidson, D. (1984). 'First Person Authority,' *dialectica*, 38/2-3: 101-111.

------- (1987). 'Knowing One's Own Mind,' *Proceedings and Addresses of the American Philosophical Association*, 60/3: 441-58.

Descartes, R. (1641/1996). *Meditations on First Philosophy* (ed. & trans. Cottingham). Cambridge, UK: Cambridge University Press.

Gendler, T.S. (2007). 'Self-Deception as Pretense,' *Philosophical Perspectives*, 21: 231-58.

------- (2008a). 'Alief and Belief,' *Journal of Philosophy*, 105/10: 634-63.

------- (2008b). 'Alief in Action (and Reaction),' *Mind and Language*, 23/5: 552-85.

------- (2011). 'On The Epistemic Costs of Implicit Bias,' *Philosophical Studies*, 156: 33-63.

Geuss, R. (2008). *Philosophy and Real Politics.* Princeton: Princeton University Press.
Greco, D. (forthcoming). 'Iteration and Fragmentation,' *Philosophy and Phenomenological Research*. Page references are to the draft manuscript.
Habermas, J. (1987). *The Philosophical Discourse of Modernity: Twelve Lectures* (trans. Lawrence). Cambridge, MA: MIT Press.
Heil, J. (1988). 'Privileged Access,' *Mind*, 97/386: 238-51.
Hume, D. (1739/2000). *A Treatise of Human Nature* (ed. Norton & Norton). Oxford: Oxford University Press.
Kamolnick, P. (1998). 'Visions of Social Justice in Marx: An Assessment of Recent Debates in Normative Philosophy,' in Panasiuk & Nowak (eds.), *Marx's Theories Today*. Amsterdam: Rodopi.
King, M. (2009). 'Clarifying the Foucault-Habermas debate: Morality, ethics, and 'normative foundations',' *Philosophy & Social Criticism*, 35/3: 287-314.
Kolodny, N. (1996). 'The ethics of cryptonormativism: A defense of Foucault's evasions,' *Philosophy & Social Criticism*, 22/5: 63-84.
Leader Maynard, J. & Worsnip, A. (ms.). 'Normativity and cryptonormativity in political realism,' draft manuscript, University of Oxford/Yale University.
Mackie, J.L. (1977). *Ethics: Inventing Right and Wrong.* Harmondsworth: Penguin.
Maes, J., Tarnai, C. & Schuster, J. (2012). 'About Is and Ought in Research on Belief in a Just World: The Janus-Faced Just World Motivation,' in Kals & Maes (eds.), *Justice and Conflicts: Theoretical and Empirical Contributions*. Heidelberg: Springer.
Mele, A. (1997). 'Real Self-Deception,' *Behavioral and Brain Sciences*, 20/1: 91-102.
Moore, C. (2013). 'This obsession with Ethics is one of the great curses of our time,' *The Telegraph*, 22nd November.
Phillips, J. & Worsnip, A. (ms.). 'Motivating internalism,' draft manuscript, Yale University.
Philp, M. (2010). 'What is to be done? Political theory and political realism,' *European Journal of Political Theory*', 9/4: 466-484.
Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.
Schwitzgebel, E. (2010). 'Acting Contrary to our Professed Beliefs, or, The Gulf Between Occurrent Judgment and Dispositional Belief,' *Pacific Philosophical Quarterly*, 91: 531-553.
------- (2011). *Perplexities of Consciousness*. Cambridge, MA: MIT Press.
Shoemaker, S. (2009). 'Self-Intimation and Second-Order Belief,' *Erkenntnis*, 71: 35-51.
Strawson, P. (1962). 'Freedom and Resentment,' *Proceedings of the British Academy*, 48: 1-25.
Sturgeon, S. (2007). 'Normative Judgement,' *Philosophical Perspectives*, 21: 569-87.
Svavarsdóttir, S. (1999). 'Moral Cognitivism and Motivation,' *Philosophical Review*, 108/2: 161-219.
Williams, B. (2005). *In the Beginning Was the Deed: Realism and Moralism in Political Argument* (ed. Hawthorn). Princeton: Princeton University Press.
Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.
Wittgenstein, L. (1953). *Philosophical Investigations* (trans. Anscombe). Oxford: Blackwell.

Tamar Schapiro
Stanford University

Tamar Schapiro is an Associate Professor of Philosophy at Stanford University. She received her Ph. D. in Philosophy from Harvard University in 1997 and was a member of the Harvard Society of Fellows from 1997 to 2000. Her interests include the nature of the passions and inclinations and their role in practical reasoning, the structure of agency, the role of ideal concepts in moral theory, and Kantian nonideal theory.

"What Are Theories Of Desire Theories Of?"

Commentator: Anthony Laden, University of Illinois - Chicago

The aim of this paper is to clear ground so that I bring a puzzle into view. My topic, at the most general level, is human motivation. To be motivated is not simply to be moved. It is to be self-moved. What is that? My more specific topic is the role of desire in motivation. What is it to have a desire, and how does desire contribute to self-movement? Now I actually believe this question is ill posed, unless we say more about how "desire" is to be understood. But I also believe this question becomes deeply interesting, from a philosophical standpoint, when we take "desire" to refer to the motives Thomas Nagel called "unmotivated desires."[160] Unmotivated desires are motives that arise in us spontaneously or automatically, without our having arrived at them as conclusions of practical deliberation. They are motives that we might call "passions," to mark the fact that we are, in this respect, passive in relation to them.[161] The question then becomes: how is the concept of a passion even coherent? The concept of a passion combines two ideas: that of motivation, and that of passivity. But motivation is self-movement. How can we be passive with respect to our own self-movement? If my passion is my motive, then it cannot be like an ocean tide, something that simply carries me along. It must, somehow, be me moving myself. But then it looks as though I must be active in relation to it. So how is passion even possible? Call this "the paradox of passion."

Some of the ideas here appeared in the first half of a paper I presented at the Pacific Division Meeting of the American Philosophical Association in March, 2013. I developed the paper further for presentations at the Dartmouth Summer Workshop on Ethics and Practical Reason in July, 2013, at the Berkeley Law and Philosophy Workshop (BAFFLE) in November, 2013, and at the University of Texas at Austin, Analytic Philosophy Symposium in December, 2013. I am extremely grateful for comments I received from those who participated in each of those discussions.

[160] Thomas Nagel, *The Possibility of Altruism*, p. 29.

[161] In Section III I will say more about the sense in which we are passive in relation to such desires. For now I leave it open how our passivity in relation to them is similar to or different from our passivity in relation to external forces, like ocean tides, or internal forces, like allergic sensitivities. I leave it open, because my aim in this paper is to argue that this is precisely the question that a theory of desire in the sense of "passion" has to address.

The reason I need to clear ground is that the recent philosophical literature on desire makes it hard to even acknowledge this puzzle, much less solve it. Contemporary philosophers, many of whom otherwise disagree with one another about a broad range of issues, have converged on a common conception of desire that I will call, following Talbot Brewer, "the evaluative outlook conception of desire."[162] I will argue that the concept of "desire" at the center of the evaluative outlook conception is deeply ambiguous, and that the persistence of the ambiguity makes it hard to see and confront the paradox of passion. Following G. F. Schueler, I will distinguish between two senses of "desire" that are often conflated in the philosophical literature.[163] I will call these the "placeholder" and the "substantive" senses of desire. Interpreted as a theory of desire in the placeholder sense, the evaluative outlook conception is, I claim, a relatively uncontroversial position that a broad range of philosophers can agree on. But so interpreted, it does not even purport to be a theory of the motives with respect to which we are distinctively passive. Interpreted as a theory of desire in the substantive sense, the evaluative outlook conception does purport to be a theory of such motives, but it fails to acknowledge, much less resolve, the paradox of passion.

I proceed here as follows. In section I, I lay out the main features of the evaluative outlook conception, and I suggest a way to understand the primary philosophical motivation behind it. In section II, the main section of the paper, I identify a tension internal to this conception. I claim that the philosophical worry that motivates the evaluative outlook conception in the first place reappears in a more local form within the conception itself, as a version of the paradox of passion. I also claim that proponents of the evaluative outlook conception do not fully recognize and confront this tension. This is so, I argue, because they tacitly rely on a mistaken assumption about the relation between the two senses of "desire." In Section III, I consider several ways in which proponents of the evaluative outlook conception might be interpreted as having at least implicitly acknowledged and responded to the tension I have identified. I claim these responses are unsatisfying, either because they simply push the problem back to another level, or because they make our relation to our feelings look too much like our relation to our actions.[164]

I. The Evaluative Outlook Conception of Desire

The evaluative outlook conception of desire is best understood as a reaction against what might be called a "brute force" view of desire. Harry Frankfurt articulates a version of a brute force view in this and other passages:

> However imposing or intense the motivational *power* that the passions mobilize may be, the passions have no inherent motivational *authority*. In fact, the passions do not really

---

[162] Talbot Brewer, *The Retrieval of Ethics*, p. 25.
[163] G.F. Schueler, *Desire: Its Role in Practical Reason and the Explanation of Action*, p. 29.
[164] I make these claims in "The Nature of Inclination." My aim here is to explain my dissatisfaction with contemporary theories of desire more clearly and persuasively than I did in that article.

> make any *claims* upon us at all. Considered strictly in themselves, apart from whatever additional impetus or facilitation we ourselves may provide by acceding to them, their effectiveness in moving us is entirely a matter of *sheer brute force*. There is nothing in them other than the magnitude of this force that requires us, or that even encourages us, to act as they command.[165]

The evaluative outlook theorist rejects two features of this brute force view. The first is the claim that desires influence us by sheer power, rather than by engaging our practical thinking. To have a desire, the evaluative outlook theorist argues, is not to be pushed or pulled by a psychological force. It is to "see" features of our circumstances as directly appealing to us as practical thinkers. To have a desire for a drink of water, he contends, is not simply to have a "blind urge" to drink a certain substance, but rather to be aware of the unpleasant dryness in one's throat, and the thirst-quenching properties of water, as together calling for, or counting in favor of, or as making appropriate, water-drinking.

The second feature of the brute force view that the evaluative outlook theorist rejects is its characterization of desire as something that influences or impinges on us, such that we have to respond or relate to *it*, rather than to the world. Positing the desire as a substantial thing that lies between us and our circumstances, the evaluative outlook theorist claims, is positing one entity and one thought, too many.[166] If the desire is any sort of thing at all, it is a perspective on the world. The desire operates in the "background," shaping our outlook such that the world appears to us in practically salient terms.[167]

Convergence on the evaluative outlook view is so broad as to make strange bedfellows. Proponents include T.M. Scanlon, a rationalist, Simon Blackburn, a Humean, Talbot Brewer, an Aristotelian, Stephen Darwall, a Kantian, and Sergio Tenenbaum, a defender of some features of scholasticism.[168] Obviously, there are important differences between these views. For one thing,

---

[165] Frankfurt, "Autonomy, Necessity, and Love," in *Necessity, Volition, and Love*, p. 137. See also Frankfurt "Reply to T.M. Scanlon," in Buss and Overton, *Contours of Agency*, p. 184: "Our most elementary desires come to us as urges or impulses; we are moved by them, but they do not as such affect our thinking at all. They are merely psychic raw material. A desire provides us not with a reason, but with a problem – the problem of how to respond to it. Impulses and urges have power, but they have no authority. They move us more or less strongly, but they make no claims on us."

[166] See, especially Simon Blackburn, *Ruling Passions*, pp. 254-55.

[167] Philip Pettit and Michael Smith, "Backgrounding Desire," p. 574

[168] Scanlon, *What We Owe To Each Other*, esp. pp. 37-41, and "Reasons and Passions," in Buss and Overton, *Contours of Agency*; Blackburn, *Ruling Passions*, esp. pp. 250-56; Brewer, *The Retrieval of Ethics*, esp. pp. 24-34; Darwall, *The Second-Person Standpoint*, esp. pp. 5, 160, and 216; Tenenbaum, *Appearances of the Good,* esp. pp. 21-51. Another contemporary philosopher who appears to endorse the essentials of the evaluative outlook view is Mark Schroeder, *Slaves of the Passions*, p. 157, n. 11. Important predecessors are G.E.M.

their targets are not exactly the same. Scanlon argues against a kind of neo-Humeanism that takes desires to be brute psychological occurrences that, in themselves, count as reasons for action. Blackburn argues against the Kantian "mistake" of thinking that we can take up a desire-free perspective on our desires. Brewer argues against what he calls "propositionalism" about desire, the view that desires have purely descriptive (i.e., nonevaluative, nonnormative) content. Darwall argues against a "naïve first-personal" account of moral reasons and moral obligation. Tenenbaum argues against a "separatist" conception of reasons, according to which justifying reasons are completely independent of motivating reasons.

In addition to differences in their targets, there are differences in their positive accounts of the type of thinking or awareness involved in having a desire. Scanlon maintains that having a desire involves seeing considerations as reasons for action, which amounts to seeing features of our circumstances as counting in favor of acting in certain ways. Brewer claims having a desire involves thoughts of the goodness of the object of desire. Blackburn seems to hold that having a desire involves awareness of certain objects as attractive or aversive, or perhaps simply as appropriate or inappropriate, in ways that reflect our particular, context-dependent needs, tastes, and interests.

Do the differences matter? I think it is important to notice that the main differences between the positive accounts can plausibly be explained as reflections of prior differences in their respective conceptions of action, rather than as differences in their conceptions of desire *per se*. Scanlon holds that desiring involves thinking about reasons, and he also thinks acting involves thinking about reasons. Tenenbaum and Brewer believe desiring involves thoughts of goodness, and they also think action involves thoughts of goodness. Blackburn thinks desiring involves thinking about the attractiveness or appropriateness of a certain action, and he thinks action involves the same. Why the correspondence in each case?

The philosophical motivation for such a correspondence is clear. If "having a desire" were wholly unlike engaging in action – if, for example, having a desire were analogous to being pushed or pulled by an external force, and engaging in action were a matter of taking considerations as reasons – then it would be very hard to see how having a desire could, in principle, make a motivational contribution to action or figure in its explanation. If my desire pushes me around like an ocean tide, then it is hard to see how its effects can, in principle, count as my actions, unless action is just a way of being pushed around.

I believe this worry is one of the main philosophical considerations that drives these very different philosophers to reject the brute force view in favor of the evaluative outlook conception. Evidence of this is the fact that nearly all evaluative outlook theorists cite Warren

---

Anscombe, *Intention,* pp. 70ff., Warren Quinn, "Putting Rationality in its Place," in *Morality and Action*, and Dennis Stampe, "The Authority of Desire."

Quinn's seminal article, "Putting Rationality in its Place," as inspiration. As I read him, Quinn is motivated by just this concern. He presents his argument as a development of Elizabeth Anscombe's remark that one could not "want" a saucer of mud unless one saw something good or worthwhile about getting it.[169] The position he opposes is a "noncognitivist" or "subjectivist" view characterized by two claims: 1) actions are explained by functional states that dispose the agent to behave in certain ways, and 2) the same functional states that explain a given action also rationalize the action – they constitute or otherwise generate reasons for the agent to do what they dispose him to do. Quinn's official claim is that such a functional state cannot in principle rationalize action in the way the subjectivist claims it can. To make the argument, Quinn asks us to imagine that he is in a functional state that disposes him to turn on every radio he sees. When he sees a radio that is turned off, he is reliably, causally disposed to turn it on. There is nothing more to this functional state. In particular, this state can be characterized without attributing to him any purportedly rationalizing thoughts, thoughts about why it makes sense to turn on every radio (e.g. in order to hear music or the news). But it is not in any way lacking, simply *qua* functional state. It takes certain conditions as input and it reliably and predictably produces behavior as output. Quinn remarks: "I cannot see how this bizarre functional state in itself gives me even *prima facie* reason to turn on radios, even those I can see to be available for cost-free on-turning."[170] The functional state may explain his behavior in some causal sense, but this explanation fails to amount to a rationalizing explanation.

Why does the state fail to rationalize? It fails, according to Quinn, because the explanation it provides does not characterize the action as something that even purports to make sense to the agent, from his own point of view, as something answerable to standards of rationality in even a thin sense. What is missing is the agent's own recognition of, or commitment to, a standard of action. Quinn has in mind a standard involving evaluative concepts, concepts of good and bad. But the argument goes through even if we conceive them as involving the normative concept of a reason for action, or perhaps even a primitively evaluative concept of the attractive or the appropriate. "I cannot see how," Quinn remarks, "in the absence of objective prior standards for evaluating ends or actions as good or bad in themselves, a state disposing one to act can be any more rationally criticizable than a state disposing one to sneeze."[171] Quinn's point, I take it, is not really that such standards of action must exist, but rather that the agent has to take his action to be subject to such standards. So interpreted, the claim is that unless the agent sees himself as subject to standards of action, whether normative or evaluative, his action can neither make sense nor fail to make sense as an intelligent, self-guided response to the way he conceives his circumstances. And if it can neither make sense nor fail to make sense in this way, then it cannot be the agent's action. It can only be something that happens through or despite him, on par with a sneeze.

I believe most, if not all, evaluative outlook theorists are sensitive to some version of Quinn's worry. They are concerned that the brute force view shows how we could be overpowered by our

---

169 Anscombe, *Intention,* pp. 70ff.
170 Quinn, "Putting Rationality in its Place," p. 237.
171 Quinn, "Putting Rationality in its Place," p. 244.

desires, but not how we could act on them. It shows how our desires can move us, but not how they can motivate us. For now, I want to grant that this worry is coherent and legitimate. But as I will show in the next section, simply endorsing the evaluative outlook conception is not enough to put Quinn's worry to rest. The worry reappears in a more local form when we press the evaluative outlook theorist to disambiguate the concept of desire he purports to explain. When "desire" is understood in a familiar sense, to refer to a type of motive with respect to which we are distinctively passive (along the lines of Nagel's "unmotivated desires"), the more local version of Quinn's worry arises. Even if such desires are perspectives on the world, to the extent that we are distinctively passive in relation to them, it makes sense to ask how they can motivate us, instead of just moving us. To be moved by a perspective, as such, is not necessarily to move oneself. The perspective has to be one's own.

II. Tensions within the Evaluative Outlook Conception

As G.F. Schueler has very helpfully shown, there is an ambiguity in the way we use the concept desire, both in ordinary life and in the philosophical literature.[172] There is one sense of "desire" or "want," such that whenever you act (where the idea of action implies that it was in some sense free, intentional, voluntary, etc.), we can say you had a "desire" to do what you did. In this sense of "desire," it is logically impossible to do something without "having a desire" to do it. To attribute a "desire" in this sense is just to attribute motivation to the agent, as the conceptual correlate of action. But there is another sense of "desire" or "want," that allows for the possibility of doing something without having a desire to do it. When you take out the garbage even though you do not feel like taking out the garbage, you do something even though you have no desire, in the second sense, to do it. You lack a certain kind of motivation. But we can still attribute to you a desire to take out the garbage, in the first sense. Nothing forced you to take out the garbage. You made yourself do it, but you did that of your own accord.

It is important to notice that the first sense of desire is actually a "placeholder," a dummy concept.[173] It is true as a conceptual matter that every action, insofar as it is attributable to the agent and not to a force external to the agent, is motivated by the agent's "desire" in the first sense. But "desire" here simply takes the place of an explanation of how the agent, rather than something external to the agent, is its source. Positing a desire in this sense does not commit us to any determinate account of what this sort of explanation looks like. Nor, importantly, does it commit us to the existence of some more specific motivational element that plays a determinate role in such an explanation.

---

[172] Schueler, *Desire: Its Role in Practical Reason and the Explanation of Action*, p. 29.

[173] Schueler, pp. 34-35. Whereas he uses the term "pro-attitude" to refer to desire in this sense, I prefer to use, "desire in the placeholder sense." I think the use of "pro-attitude" presupposes that the explanation of an action takes the form of a reference to a mental state that causes or leads to the action. I do not want to commit to that conception of action explanation here.

By contrast, the second sense of "desire," according to which it is logically possible to do something without having a desire to do so, cannot just be a placeholder for an explanation of action. You had no desire to take out the garbage, so whatever explains what you did makes reference to something other than this absent desire. And presumably in a different case, where you take out the garbage because you feel like taking out the garbage (say, you are enjoying purging your closet), that desire would play a determinate role in the explanation what you do. I will call the second sense of desire, the "substantive" sense, because that concept does pick out something that plays a distinct, determinate role in the explanation of action.[174]

Which concept of desire is the evaluative outlook conception a conception of? This turns out to be a difficult question. Since most versions of the view take Quinn's article as inspiration, they tend to inherit his concept. Quinn, in turn, inherits his concept from Anscombe, and from the neo-Humean functionalist view he opposes. I believe it makes the most sense to read all three as using "desire" in the placeholder sense. Indeed, the logic of Quinn's argument makes sense on this reading. For Quinn is making a point about what it takes to explain action *qua* action. The negative part of his claim is that a functional state, as such, cannot explain action because it represents the agent as simply being moved from without, rather than as guiding himself from within. The positive part of the claim is that any explanation of action has to characterize the agent as guiding himself according to standards of action that he takes to be genuinely binding on him. Read in this way, Quinn's argument simply claims to identify a necessary (if not sufficient) condition on action explanation. It does not purport to give an explanation of desire as a determinate element in that explanation, let alone one with respect to which we are distinctively passive.

But defenders of the evaluative outlook conception of desire routinely appeal to Quinn's argument for support. To that extent, we should ask whether these inheritors of Quinn are claiming to defend a conception of desire in the placeholder sense or in the substantive sense. The worry is that they illegitimately slide from a conception of desire as action explanation itself, to a conception of desire as a determinate element in the explanation of action.

Talbot Brewer's theory, though unique in some respects, is illustrative of a tendency common to many versions of the evaluative outlook approach. The tendency is to rely on the placeholder sense of desire when making the core argument for the evaluative outlook view, and then to claim that the same view accounts for the substantive sense of desire, modulo certain minor qualifications. The substantive sense of desire is thus treated as if it were a special case of the placeholder sense of desire. Later I will argue that this is problematic. The relation between the placeholder sense of desire and the substantive sense is not that of genus to species. But first let me simply make clearer the tendency I find problematic.

---

[174] Schueler calls this "desire proper" (p. 35). I prefer "desire in the substantive sense" because it contrasts more clearly with "desire in the placeholder sense."

Brewer's main argument for the evaluative outlook approach is exactly Quinn's. Brewer writes:

> We cannot provide a rationalizing explanation of an episode of agency simply by tracing it to some class of performances that the agent is disposed to produce, since the agent might see no more point in these performances than in obsessions or nervous tics. This is what lies behind Anscombe's claim that desiring requires that the desirer see something desirable in that which is desired. Ordinary usage of the term 'desire' is perhaps ample enough to encompass merely obsessive urges. However, tracing behavior to such a bare urge . . . does not suffice to make the behavior intelligible as something the person saw fit to choose. No conception of desire can play a central role in rationalizing explanations of action – that is, in making behavior intelligible as action – unless it takes desiring to consist at least partly in appearances of desirability.[175]

Notice that what Quinn had referred to as a "functional state," Brewer (following Anscombe) refers to as a "desire." But the role of the concept in the argument is the same. It is a placeholder for the explanation of the action in question. Let us agree with Anscombe, Quinn, and Brewer that this explanation must reveal the agent to be guiding himself according to some conception of what is valuable or reasonable or attractive. The question is whether Brewer sees this as a claim not only about action explanation as such, but also about desire in the substantive sense, as a determinate element that plays a distinctive role in the explanation of action.

At the end of the chapter in which this main argument appears, Brewer does acknowledge the distinctiveness of desire in the substantive sense. The issue comes up almost as an afterthought:

> I conclude, then, that desires are best understood as consisting not just partly but wholly in appearances of reasons or values. It must be noted, however, that not just any such appearance will constitute a desire. It is an article of common sense that we sometimes decide that we have good reason to do something, and proceed to do it, even though we have no desire whatsoever to do it. It can hardly be said in such cases that it in no way appears to us that there is reason to do what we've chosen to do. We must have seen and acted on *some* reason for doing it – otherwise there would be no sense in thinking of it as something we've done. Such an appearance, however, need not count as a desire, since it might come into focus only with sustained deliberative effort. Desires are appearances with respect to which we are in some significant measure passive. Their occurrence is not wholly dependent upon our active efforts to bring into view the appearances of goodness in which they consist, and their persistence and vividness does not depend entirely on our deliberate efforts to discern the putative goods they call to our attention.[176]

Is Brewer claiming to have offered a conception of desire in the placeholder sense, or in the substantive sense? On the one hand, the Anscombe/Quinn line of reasoning leads to a conclusion about action explanation as such. The fact that we sometimes act without having a desire in the substantive sense is entirely compatible with the central claim that whenever we act, we see something good, or reasonable, or attractive, or appropriate in doing what we do. When you take

---

[175] Brewer, *The Retrieval of Ethics,* p. 24. Later, when referring back to this main argument, Brewer cites Quinn explicitly (p. 55, n. 11).
[176] Brewer, *The Retrieval of Ethics,* p. 34.

out the garbage even though you don't feel like taking out the garbage, you are still guided by your conception of what makes taking out the garbage worthwhile. The question is whether this claim tells us anything about the desire that you happen to lack in this case – namely a desire, in the substantive sense, to take out the garbage. Since you do not have this desire, it presumably plays no role in the explanation of your taking out the garbage. Does the Anscombe/Quinn line of reasoning have any implications for the question what you lack in this case? Does Brewer think it does?

In the passage above Brewer seems to be saying that whereas all desires, in some general sense, are evaluative outlooks, the kind of desire you lack when you do not feel like taking out the garbage has a further differentiating property. It is an evaluative outlook with respect to which you are in some distinctive respect passive. It is a perspective on the good, or on reasons, that you simply find yourself occupying, whereas other desires are perspectives that you occupy because you make a deliberate effort to do so. Brewer does not elaborate. But we can imagine how his account would inform our understanding of the case at hand. Suppose you are aware that you ought to take out the garbage now, but you have no "desire" in the substantive sense to do so. What you want to do, in the substantive sense, is to curl up on the couch. No matter how you act, whether you take out the garbage now or curl up on the couch now, the explanation of your action will refer to your occupying a perspective from which you are aware of the goodness or reasonableness or attractiveness or appropriateness of what you are doing. If pushed to give a more fine-grained explanation that allows us to see the difference between these two cases, what we should say is this. In the scenario where you take out the garbage now, the perspective that guides you is one you actively inhabit through "sustained deliberative effort." In the scenario where you curl up on the couch now, the perspective that guides you is one you happen to find yourself occupying.

But the explanation in the second case raises a local version of the same worry that motivated the Anscombe/Quinn line of thought in the first place. If the desire to curl up on the couch is a perspective you simply find yourself occupying, rather than one you actively inhabit, how is that perspective your own, and how does being guided by it count as you guiding yourself? On Brewer's view, your thoughts of how good it would be to curl up on the couch are not thoughts you have arrived at actively, by deliberating. But then how are you related them? Are you in any sense doing the thinking? If not, then in being guided by these thoughts, why haven't you simply been hijacked or possessed? Granted, the model of alienation here is not that of being pushed or pulled by a "blind" force. It is that of being infiltrated by something more agential, a way of seeing and responding to the world. But it is still a model of alienation.

I am not claiming Brewer could not develop his view in such a way as to resolve this problem. My point is that he does not acknowledge it as a problem. He does not recognize that since what is distinctive about desire in the substantive sense is our passivity in relation to it, desire in that sense cannot explain action without raising a local version of the general problem that motivated the evaluative outlook view in the first place. I believe other versions of the evaluative outlook conception are insufficiently developed in exactly the same respect. But to make that case, I want

to turn to T.M. Scanlon's theory, which is arguably the most detailed contemporary articulation of the evaluative outlook conception.

Like Brewer, Scanlon cites Quinn's radio man example approvingly in order to motivate his own version of the evaluative outlook conception of desire.[177] Later, in a direct reply to Frankfurt, Scanlon worries explicitly that if desires are merely brute forces, then the agent who is motivated by a desire "is thus not acting on a reason at all, but only being overpowered by an impulse."[178] On Scanlon's view, by contrast, desiring involves normative thinking:

> When I feel a desire for a piece (or a second piece) of rich chocolate cake, its delicious taste and the pleasure it would give me seem to me to be reasons for eating it. When I feel a desire for revenge against my rival, the fact that something I could do would cause him embarrassment strikes me as a reason to do it.[179]

Scanlon believes this account of desire does a better job of showing us how desire leads to actions done on the basis of reasons, while better capturing the phenomenology of desiring. When you act on your desire to eat the piece of chocolate cake, you are acting on what you were already taking to be reasons to eat it.

Which concept of desire is Scanlon attempting to capture? Scanlon is well aware of the difference between desire in the placeholder sense and desire in the substantive sense. And while he certainly holds that his account is true of desire in the placeholder sense, his primary aim is to show that it is true of desire in the substantive sense as well. In other words, he realizes that there is a trivial sense of "desire" according to which, every time a person acts, he does what he "wants" to do. And he also believes that every time a person acts, the explanation of that action makes reference to the agent's having taken considerations as reasons. So he would not deny that his account is true of desire in the placeholder sense. But he also recognizes that someone like Frankfurt might argue that his account fails to capture what is distinctive about desire in the substantive sense. Notably, we do not generate the desire to eat the cake by first arriving at a judgment that there are reasons to eat it. The desire simply arises in us, spontaneously. Moreover, the desire to eat the chocolate cake has a motivational power that seems to be independent of what we take to be the strength of the justification for eating it. For we often want to do what we judge we should not do, all things considered. The challenge for a view like Scanlon's is to do justice to these features of desire in the substantive sense.

---

[177] *What We Owe To Each Other*, pp. 38 and 43. I should note that although Scanlon refers to Quinn's argument approvingly, he does not, in these passages, directly appeal to Quinn's considerations about action explanation as grounds for embracing his version of the evaluative outlook view. Instead he seems to appeal directly to the phenomenology of desiring. The considerations about action explanation are more clearly operative in his reply to Frankfurt, "Reasons and Passions," in Buss and Overton, *Contours of Agency*, pp. 165-183.

[178] Scanlon, "Reasons and Passions," in Buss and Overton, *Contours of Agency*, p. 177.

[179] Scanlon, "Reasons and Passions," p. 178.

Scanlon explicitly addresses this challenge when he writes:

> I might seem to be saying here that there is no such thing as an unmotivated desire. Taken in Nagel's sense this would entail that all desires arise from prior evaluative judgments of some kind, a claim that seems clearly false. What I am claiming, however, is not that all desires arise from prior judgments but rather than having what is generally called a desire involves having a tendency to see something as a reason. Even if this is true, however, this is not all that desire involves . . . I might see something good about drinking a glass of foul-tasting medicine, but would not therefore be said to have a desire to do so . . . Reflection on the differences between these cases leads me to what I will call the idea of desire in the directed-attention sense. A person has a desire in the directed-attention sense that P if the thought of P keeps occurring to him or her in a favorable light, that is to say, if the person's attention is directed insistently toward considerations that present themselves as counting in favor of P.[180]

Scanlon is not claiming that beliefs about reasons or goodness are motivationally inert without something further, called "desire." He rejects this Humean view. Rather, his point is that in some cases we act without having a desire in the substantive sense. He believes his account of desire, in a very general form, does tell us something about even those cases. Even when we drink the foul-tasting medicine without "wanting to" in the ordinary sense, it is the case that we are taking considerations, e.g. that the medicine will alleviate our symptoms, as a reason to do what we do. So in that respect, Scanlon is offering an account of desire in the placeholder sense. But he is more interested in offering an account of desire in the substantive sense. And like Brewer, he assumes that the way to do this is to add differentia to his general conception of desire. That is, he tacitly relies on an assumption that desire in the substantive sense is a species of desire in the placeholder sense.

We can ask whether the differentia he adds are the right ones, or we can ask the more fundamental question whether adding differentia is the way to meet the challenge. As to the first of these questions, it is not entirely clear, as a phenomenological matter, that what is distinctive about desires in the substantive sense is their insistence. Suppose you are procrastinating. You are relaxing on the couch instead of taking out the garbage, and you are aware of yourself as acting against your better judgment. In these cases the guilty conscience can be very insistent. That is why doing what you feel like doing as a way of procrastinating is so much more painful than doing what you feel like doing when you are not procrastinating. But we would not say that the insistent prickings of conscience are desires in the substantive sense.[181]

---

[180] Scanlon, *What We Owe To Each Other*, p. 39.

[181] Scanlon remarks that desires in the directed attention sense can take any object, including "to do the right thing." (*What We Owe To Each Other*, p. 39) So a person with an active conscience "has a strong desire to do the right thing." But it would still be odd to say that the procrastinator is pained by his insistent conscientious desire, if we are trying to capture what is distinctive about desire in the substantive rather than the placeholder sense.

Even if it is unclear that the phenomenological mark of desire in the substantive sense is insistence, we can consider Scanlon's apparently independent claim that what distinguishes desires in the substantive sense is the way in which they arise.[182] They track Nagel's notion of unmotivated desires because they are thoughts about reasons that simply "assail us, unbidden," instead of being conclusions we arrive at through active deliberation.[183] Here the differentia is the same one Brewer notes, namely our passive relation to desire in the substantive sense. But as we saw in connection with Brewer's account, this move raises the worry about how desire in this sense is supposed to lead to action. If certain thoughts about reasons simply arise in you and cause you to act in accordance with them, why haven't you been hijacked or possessed? If desires in the "directed-attention sense" indeed "direct" your attention insistently towards their objects, who or what is doing the directing? How is this source both *you* and someone (or something) with respect to whom (or to which) you are distinctively passive? And to whom or to what are we to attribute the actions that issue from such desires?

As with Brewer, the worry that leads Scanlon to reject a brute force view of desire in the placeholder sense arises in a local form when he is pushed to explain how desire in the substantive sense is a distinct form of motivation. And like Brewer, Scanlon does not acknowledge this as a problem. He does not ask how desires in the directed-attention sense can play a role in action explanation, given that they simply assail us. He does argue, in the next section of the chapter, that having a desire in this sense almost never gives us a reason to act in accordance with it. But there he is using the notion of a "reason" in a stronger sense than is required by Quinn's notion of a rationalizing explanation. Granted, the fact that you wanted to curl up on the couch is not in itself a conclusive reason (or maybe not even a reason) to for you to do so now, instead of taking out the garbage now. But Quinn's worry about dispositional accounts is a worry about a less controversial aspect of a neo-Humean view. It is not a worry about how having a desire can justify an action. It is a worry about how having a desire can lead to an action.

To recap: Brewer and Scanlon start out doing action theory. They are responsive to Quinn's worry that actions, as such, cannot be explained as the effects of brute forces. In response to this, they endorse the Quinn's positive view that action explanation involves reference to an evaluative outlook. They put this forth as a "general" conception of "desire," where what they are referring to is desire in the placeholder sense. They then note that this conception of desire is not extensionally equivalent to our "ordinary" notion of desire, that of desire in the substantive sense. In order to develop an account that does track this ordinary notion, they add a qualification

---

[182] I actually think there is a more generous reading of Scanlon on this point. By "insistence," I think he may be referring to the immediacy with which desires in the substantive sense present their objects to us in practically salient terms. But even so, we need more than a description of this feature. We need an explanation of it.

[183] Scanlon, *What We Owe To Each Other,* p. 39, referring to Nagel, *The Possibility of Altruism*, p. 29.

to the original account. A desire in the substantive sense is an evaluative outlook with respect to which we are passive, in the sense that the outlook is not a conclusion we have arrived at through active deliberation. But they do not take on the burden associated with this claim, which is to confront the paradox of passion. Grant that some evaluative outlooks simply "assail us, unbidden." How is our passivity in this respect compatible with the claim that when we are moved by these outlooks, we count as moving ourselves?

If what Scanlon, Brewer, and other evaluative outlook theorists are offering is a theory of desire in the substantive sense, then this is *the* central question they have to address. Resolving the paradox of passion is, I claim, the main task of such a theory. If the theory does nothing more than note the fact that we are distinctively passive in relation to desires in the substantive sense, then all it does is identify the relation that requires philosophical explanation.

Granted, the evaluative outlook theory is presented as doing more than this. It purportedly shows us that desires, *in general*, are, or involve, evaluative outlooks. In that case we should assess its merits not as a theory of desire in the substantive sense, but as a theory of desire in some more general sense. But what is this more general sense? As I noted earlier, Quinn's main point, which the evaluative outlook theorists reiterate, is about "desire" only in the placeholder sense. "Desire" in this sense is a placeholder for action explanation as such. The radio man argument purports to show that any explanation of action *qua* action has to represent the agent as seeing a point in his action, such that he undertook it. Granted, this makes a lot of sense. Assume that whatever else human action is, it is at least an agent's purposive response to his representation of the circumstances.[184] Then it is intuitive that any explanation of action has to appeal to the agent's way of looking at the world in practically salient terms. It seems relatively uncontroversial that an evaluative outlook has to play some role in the explanation of any action. But to call this a theory of "desire," and then on this basis to suggest that it helps us to understand what is distinctive about desire in the substantive sense, is a mistake. Desire in the placeholder sense simply is not a genus of which desire in the substantive sense is a species. Desire in the placeholder sense is a placeholder for whatever it takes to explain an action. Desire in the substantive sense is a distinctive element that plays a role in some explanations of action. The evaluative outlook theory is most plausible as a relatively uncontroversial claim within a larger theory of action explanation, or desire in the placeholder sense. It does not even pose the central question to be addressed by a theory of desire in the substantive sense.

III. Responses on behalf of the evaluative outlook conception

Let me now consider two replies on behalf of the evaluative outlook theorist. The first appeals to an analogy between desire, in the substantive sense, and perception. The second calls into

[184] I am stating this idea in an admittedly vague way, so that it can accommodate various more precise formulations. See, for example, Harry Frankfurt, "The Problem of Action," and Christine Korsgaard, *Self-Constitution: Agency, Identity, and Integrity*, esp. sections 5.4-5.6, pp. 93-108.

question the claim that our passivity in the face of such desires is philosophically significant, or generates an important philosophical problem. I will address these in turn.

An article that has been nearly as influential as Quinn's in the development of the evaluative outlook view is Dennis Stampe's "The Authority of Desire." There, Stampe writes:

> Desire is a kind of perception. One who wants it to be the case that *p* perceives something that makes it seem to that person as if it would be good were it to be the case that p, and seem so in the way characteristic of perception. To desire something is to be in a kind of perceptual state, in which that thing seems good…"[185]

Although Stampe uses "desire" in a way that is ambiguous between the placeholder and the substantive sense, I take it the appeal of this language comes from tacit recognition that there is an analogy between our passivity in relation to our perceptions and our passivity in relation to our desires in the substantive sense. Later evaluative outlook theorists, who likewise use "desire" ambiguously, draw the same analogy.[186] Scanlon, for example, distinguishes between "seeing" and "judging" considerations as reasons, or, equivalently, between "seemings" and "assessments."[187] Suppose you judge that you have no reason to be anxious about what others think of you, but you still find yourself wanting to act in ways that you believe will please them and arouse their approval of you. In that case, Scanlon claims, it is as if you are perceiving something to be a reason that you judge is not actually a reason. The experience differs from straightforward indecision, because you have made up your mind about what you ought to do. You already side with your judgment, as a conclusion you have arrived at through active deliberation. But you are still susceptible to the motivational force of reasons that simply "appear" to you, independent of your deliberation. The experience is like that of an optical illusion.

We can read the appeal to perception in one of two ways. On one reading, the claim is that desiring (in some unspecified sense) just is a form of perception, one that literally involves "seeing" reasons or goodness. This reading is compatible with a view that denies the distinction between practical and theoretical employments of reason. Practical judgment, on this view, is actually a species of theoretical judgment. On another reading, the claim is that desiring is analogous to perceiving. Desiring is related to practical judgment in the way that perceiving is related to theoretical judgment. In general, when evaluative outlook theorists make reference to perception, they do not state clearly which version of the appeal they are making. But I doubt they all share the same view of the relation between practical and theoretical reasoning. For my

---

185 Stampe, "The Authority of Desire," p. 359.

186 See, for example, Tenenbaum, *Appearances of the Good*, Karl Schafer, "Perception and the Rational Force of Desire," Christine Tappolet, "Emotions and the Intelligibility of Akratic Action," in Stroud and Tappolet, *Weakness of Will and Practical Irrationality*, and Jessica Moss, "Akrasia and Perceptual Illusion."

187 For "seeing" and "judging," see *What We Owe To Each Other*, pp. 39-40. Scanlon uses the terminology of "seemings" and "assessments" to mark the same distinction in "Reasons and Passions," p. 176.

purposes, it will be more helpful to read the reference to perception as an appeal to an analogy between practical and theoretical reasoning, conceived as distinct employments of reason.

The analogy between desire and perception is indeed very close. Like "desire," "perception" can be used in a placeholder or a substantive sense. In the placeholder sense, whenever an agent believes that X is P, we can say, trivially, that he perceives that X is P (or perceives X as P). But "perception" in that sense is just a stand-in for whatever explains his belief. It is not an element that plays a distinct role in that explanation. In the substantive sense, by contrast, it is possible to believe that X is P without perceiving that X is P (or perceiving X as P).

Perception in the substantive sense is something with respect to which we are distinctively passive. We do not arrive at our perceptions through deliberation. Rather, they "assail us, unbidden." This raises the question: when I believe what I see, simply because I see it, is my belief an exercise of my cognitive agency, or is it just the effect of a process that causes me to experience an impression with a certain degree of "force and vivacity"? How is our passivity in relation to our perceptions compatible with the idea that those perceptions can, at least in principle, make genuinely cognitive contributions to our theoretical judgments? The theorist of perception has to confront this question, just as the theorist of desire in the substantive sense has to confront the analogous question.[188] But the evaluative outlook theorist cannot point to this analogy as if it counts as an answer to either question.

Let me turn now to the evaluative outlook theorist's second reply, which I think is more significant. The claim here is that there really is no paradox of passion, because we simply are not completely or importantly passive in relation to the desires that "assail us, unbidden." True, we do not arrive at those desires as conclusions of deliberation. And we often feel that they are not under our full control. But, the evaluative outlook theorist argues, this does not mean they are not "ours" in the sense required for full-fledged motivation and action. In particular, this phenomenological feature of unmotivated desires should not be taken as evidence that they, or the thoughts they involve, are attributable to some source "outside" us, or even to a capacity "in" us that is external to our capacity to act on reasons. Indeed, he would argue, there is no good

[188] For a rich and historically illuminating discussion of recent debates about how perception can play an epistemological role, see Michael Friedman's, "Exorcising the Philosophical Tradition: Comments on John McDowell's *Mind and World.*" It should be clear that one of my aims is to encourage evaluative outlook theorists to think deeply about whether and in what sense there is a *practical* faculty of receptivity, and if so, how it is related to that of spontaneity. I share Friedman's misgivings about the way McDowell appropriates Kant in developing his own, idealist position on a version of this question that does not clearly distinguish between theoretical and practical receptivity.

philosophical reason to slide from the description of our felt passivity to an elaborate metaphysics of independent motivational sources or faculties.[189]

In support of this claim, the evaluative outlook theorist will point out that we are not *entirely* passive in relation to our unmotivated desires. For even those desires, like our beliefs and like our decisions, are, to use Scanlon's phrase, "judgment-sensitive attitudes."[190] To deny this, the evaluative outlook theorist maintains, amounts to assimilating unmotivated desires to morally irrelevant features of a person, like allergic sensitivity. It is to claim, for example, that a person's proneness to, say, lazy impulses (independent of whether he acts on them) is no more a reflection on him than is his sensitivity to pollen.[191] And it is to claim that our attitude towards our own lazy impulses (independent of whether we act on them) should likewise not differ from our attitudes towards the purely biological processes that affect our functioning. We might wish they were otherwise, and we might try to change them by means of medications or nonrational techniques, but it would make no sense to feel pride or shame because we have them, or to try to reason ourselves out of them.

That there is some difference between our relation to our allergic sensitivities and our relation to our unmotivated desires is entirely plausible, as is the idea that the latter have something to do with reason that the former do not. But these are not the only intuitions to be accounted for. There are also intuitions about how our relation to our unmotivated desires differs from our relation to the motives that we do arrive at as conclusions of deliberation, and how each is, ideally, judgment-sensitive. Suppose you are prone to aggressive impulses, especially when driving in traffic jams. The sheer fact of being obstructed by other cars tends to make you angry and hostile. Suppose your aggression takes the form of thoughts such as, "the fact that there are so many cars obstructing me is a reason for me to yell at other drivers." Despite these feelings, you judge that you have no reason to yell at other drivers. You recognize that the fact that there are a lot of other cars on the road may be disappointing, but it does not show that any of the other drivers has wronged you. Nor will yelling at anyone change anything for the better. Because you have arrived at these conclusions, you have a motive to refrain from acting on your hostile impulses. Nevertheless, you are still prone to feeling them.

Are your hostile impulses, in the ideal case, judgment-sensitive and attributable to you in the same way that your motive to refrain from yelling is judgment-sensitive and attributable to you? The evaluative outlook theorist denies that your unmotivated desire is external to you in any important sense. So the picture cannot be that your aggressive impulses, like powerful tides, set in motion a brute causal process with respect to which you are a spectator, a process that it is up to you to interrupt before it leads to certain effects. Rather, the picture has to be that you are both

---

[189] I take it this is Scanlon's main point in *What We Owe To Each Other*, 39-41. See esp. p. 40: "…we should not take 'desires' to be a special source of motivation, independent of our seeing things as reasons."

[190] Scanlon, *What We Owe To Each Other*, pp. 20ff.

[191] Scanlon stresses this general point in "Reasons and Passions."

the one who is doing the restraining, and the one who is restrained. So perhaps the idea is this. Your aggressive impulse to yell at the other drivers is simply you beginning to yell at them, on the basis of what you take to be reasons from the perspective of your unmotivated desire. Your feeling is your action is at a very early stage, a stage at which it does not yet have behavioral manifestations. Unless you stop yourself, you will complete it. So when you restrain yourself, you are aborting an action you have already begun to undertake.

This is a position available to the evaluative outlook theorist. But it does not come without a cost. It sits uncomfortably with the following feature of our thought and talk about feelings. You can decide to act, but you cannot decide to feel. You can decide not to yell at other drivers, and on that basis, make yourself refrain from yelling at them. But you cannot decide not to feel like yelling at them, and on that basis make yourself not have those feelings. If the appeal to decision here seems too voluntaristic, substitute the idea of becoming convinced. You can become convinced that you ought not to yell at other drivers, and on that basis, make yourself refrain from yelling. But you cannot become convinced that you ought not feel like yelling at them and, on that basis, make yourself not feel like yelling at them.

This is entirely compatible with the equally obvious fact that we can indeed cultivate and shape our feelings over time. We do this through imaginative and behavioral techniques. Convinced that you would like to be a person who does not even feel like yelling at other drivers, you can indeed change. You can do this by cultivating new habits. You can learn to breathe deeply during traffic jams, and you can use your imagination to practice "seeing" other drivers as fellow victims, rather than as persecutors. The point is that whereas we can change our actions by deciding to act differently, or by becoming convinced we should act differently, we can only change our feelings by using nonrational techniques to cultivate new ones over time.

This observation supports the idea that as agents, we are simply not related to our feelings in the same way that we are related to our actions. It likewise supports the idea that our feelings are simply not, even ideally, judgment-sensitive in the same way that our actions are. The fact that these impulses can in some way reflect on our character, whereas our allergic reactions cannot, need not imply that they reflect on us in the same way that our conduct does. If the evaluative outlook theorist wants to grant that there is a difference here, he has to make a commitment that goes beyond phenomenological description. He has to say something about where we stand in relation to our unmotivated desires, as compared with where we stand in relation to our motivated desires.

Here is a further fact that supports the idea that our feelings do not reflect on us in the same way that our actions do. Our reactive attitudes towards ourselves differ in kind, depending on whether we are critical of our own conduct or whether we are critical of our feelings. We hold ourselves directly accountable to ourselves for having acted against our better judgment. It is far less clear

that we hold ourselves directly accountable to ourselves for having had feelings that conflict with our better judgment.

The difference here bears analogy, I think, with the difference between reactive attitudes towards an adult and reactive attitudes towards a child. I have argued elsewhere that it is a fixed point in our practice towards children that we do not take them to be proper objects of direct resentment and blame.[192] We do not straightforwardly resent children when they do wrong. Instead we are disappointed in them. The response is not that of an equal but that of a superior. This is reflected in the idea that children are to be disciplined, where discipline is not exactly the same as punishment. Discipline is a more forward-looking response. The aim is to educate, train, reform, rather than to exact a debt and so to right the scales of justice. And the background assumption is that children are not fully responsible for what they do in the same, direct way that adults are responsible for what they do. Similarly, I want to claim, our response to ourselves when our feelings fail to support our better judgment is not that of an equal but that of a superior. We are disappointed in ourselves as sources of feeling, and we resolve to discipline and train ourselves better in that capacity. But we do not resent and blame ourselves as we would if we had acted badly. I take it this is simply a retrospective reflection of our prospective awareness that we can decide to act, but we cannot decide to feel.

Notice that this picture does not imply that feelings are brute pushes and pulls, any more than it implies that children are mere objects. The cultivation of feeling, like the raising of a child, is a matter of shaping an intelligent perspective on the world, and is not just a matter of instilling predictable responses to stimuli. That said, there is obviously a limit to the analogy. The children we raise are separate people, with lives of their own. Our feelings are not separate from us in this sense. Even if they start out having lives of their own, their proper role is to contribute to our lives.

Does the evaluative outlook conception recognize what I have taken here to be a fixed point in our thought and talk about feeling? Does it acknowledge that our feelings are not identifiable with us (either from our own perspective, or from that of others) in the same, direct way that our conduct is? On this point, again, the view is deeply ambiguous. Scanlon holds that desires, even in the unmotivated sense, are judgment-sensitive attitudes. In this respect, he claims, they are like beliefs and decisions. But this does not tell us whether having unsupportive feelings is *our* failure in the same sense that refusing to act on our better judgment is *our* failure. The notion of judgment-sensitivity is too coarse-grained to register this distinction. Granted, there surely is some coarse-grained sense in which both our feelings and our deliberative conclusions are "internal" to us, or are "up to us," or are motives for which we are responsible. But in order to articulate what is distinctive about unmotivated desires as specific elements in the explanation of action, the evaluative outlook approach needs to appeal to a more fine-grained conceptual vocabulary. And that vocabulary has to reflect a substantive theory of how the basic normative

[192] See my "Childhood and Personhood."

relation in which we stand to our feelings is different from the basic normative relation in which we stand to our actions.

IV. Conclusion

In this paper I have tried to undermine complacency with a predominant conception of desire, for the sake of refocusing attention on a philosophical problem. I have argued that it is not clear what the evaluative outlook theory of desire is a theory of. If it is a theory of desire in the placeholder sense, then it is at bottom a theory of action explanation. So construed, its claim is relatively uncontroversial, and falls far short of being a full theory of action explanation. The claim is simply that the agent's way of looking at the world in practically salient terms must play a part in the explanation of his action. If, on the other hand, the evaluative outlook conception is a theory of desire in the substantive sense, then it does not even go so far as to acknowledge the central problem such a theory has to answer. That problem is how we can be passive in relation to our own self-movement. I have suggested that the evaluative outlook theorist's appeal to an analogy desire and perception, though not misguided, simply pushes the question back. Moreover, his assertion that desires, like beliefs, are judgment-sensitive, is too coarse-grained to account for important differences between merely feeling like doing something and doing it. What we need is an account of our relation to our feelings, one that takes seriously our passivity with respect to them without assimilating them to external forces. Ultimately, I believe this relation is *sui generis.* We cannot fully capture it by analogy with other relations, such as that of a charioteer to his horse, or an adult to a child, or even a State to its citizens. But that is what makes it so worthy of philosophical attention.

WORKS CITED

Anscombe, Gertrud Elisabeth Margaret. *Intention*. Harvard University Press, 1957.

Blackburn, Simon. *Ruling passions*. Oxford: Clarendon Press, 1998.

Brewer, Talbot. *The retrieval of ethics*. Oxford University Press, 2009.

Buss, Sarah, and Lee Overton, eds. *Contours of agency: Essays on themes from Harry Frankfurt*. The MIT Press, 2002.

Darwall, Stephen L. *The second-person standpoint: Morality, respect, and accountability*. Harvard University Press, 2006.

Frankfurt, Harry G. "Reply to T.M. Scanlon." *Contours of agency: Essays on themes from Harry Frankfurt* (2002): 184-188.

Frankfurt, Harry G. *Necessity, volition, and love*. Cambridge: Cambridge University Press, 1999.

Frankfurt, Harry G. "Autonomy, necessity, and love." *Necessity, volition, and love* (1999): 129-141.

Frankfurt, Harry G. *The importance of what we care about: Philosophical essays*. Cambridge University Press, 1988.

Frankfurt, Harry G. "The Probem of Action." *The importance of what we care about* (1988): 69-79.

Friedman, Michael. "Exorcising the Philosophical Tradition: Comments on John McDowell's Mind and World." *The Philosophical Review* 105, no. 4 (1996): 427-467.

Nagel, Thomas. *The possibility of altruism*. Princeton University Press, 1978.

Pettit, Philip, and Michael Smith. "Backgrounding desire." *The Philosophical Review* 99, no. 4 (1990): 565-592.

Quinn, Warren. *Morality and action*. Cambridge University Press, 1993.

Quinn, Warren. "Putting rationality in its place." *Morality and action* (1993): 228-55.

Schueler, George F. *Desire: Its role in practical reason and the explanation of action*. The MIT Press, 1995.

Scanlon, Thomas M. "Reasons and Passions." *Contours of agency: Essays on themes from Harry Frankfurt* (2002): 165-183.

Scanlon, Thomas M. *What we owe to each other*. Harvard University Press, 1998.

Schapiro, Tamar. "The Nature of Inclination." *Ethics* 119, no. 2 (2009): 229-256.

Schapiro, Tamar. "Childhood and personhood." *Ariz. L. Rev.* 45 (2003): 575.

Schapiro, Tamar. "What Is a Child?" *Ethics* 109, no. 4 (1999): 715-738.

Schroeder, Mark Andrew. *Slaves of the Passions*. Oxford: Oxford University Press, 2007.

Stampe, Dennis W. "The authority of desire." *The Philosophical Review* 96, no. 3 (1987): 335-381.

Tenenbaum, Sergio. *Appearances of the good: An essay on the nature of practical reason*. Cambridge University Press, 2007.

Dylan Murray and Lara Buchak
University of California-Berkeley

Dylan Murray is a graduate student at the University of California, Berkeley. His main interests lie in moral psychology and action theory, and he has published on folk intuitions about moral responsibility and free will. Lara Buchak is assistant professor of philosophy at the University of California, Berkeley. Her main interests lie in decision, game, and rational choice theory. Her book *Risk and Rationality* concerns how risk ought to be taken into account in decision-making.

"Risk and Motivation: Why 'What To Do?' Isn't Settled By 'What Should I Do?'"
Commentator: Debbie Goldgaber, Northwestern University

Abstract: *Within philosophy of action, there are three broad views about what, in addition to beliefs, constitute an answer to "what to do?": desires (Humeanism), judgments about values/reasons (rationalism), or states of the will, like intentions (volitionalism). We argue that recent work in decision theory – risk-weighted expected utility theory (Buchak 2013) – vindicates the volitionalist. "What to do?" isn't settled by "what should I do?" Rather, rational motivation requires determining how to trade off the possibility of a good outcome against the possibility of a bad one, i.e., determining how much of a risk to take. These risk attitudes are best understood as intentions or self-governing policies to weight desires or reasons in certain ways, and are required to resolve impasses of evaluative or normative underdetermination. Far from being rare, though, or confined to Buridan's-ass-like cases, such underdetermination is in fact typical of choice under uncertainty.*

I. Introduction

Decision theory and philosophy of action both attempt to explain what it is for an ideally rational agent to answer the question *"what to do?" (w.t.d?)*. That answer is, from the agent's point of view, the conclusion of her practical deliberation, and the mental states that constitute it are the sources of reasons-explanations of her behavior.

In philosophy of action, there are three broad views about what constitutes an answer to that question (Wallace 2006) – i.e., about *what types of mental states*, in addition to beliefs or credences, determine w.t.d.: desires (the Humean), evaluative/normative[193] judgments or cognitive beliefs about what's good, valuable, or what one has most reason to do (the rationalist), or (perhaps in addition to either or both of the above) some separate states of the will, like intentions or plans, not fully reducible to any combination of desires or judgments (the volitionalist).

Rationalists claim that "what to do" is completely settled by answering the question "what should I do?"[194] Volitionalists claim otherwise, and typically appeal to *akrasia* and

[193] For expository easy, we'll use 'evaluative' and 'normative' more or less interchangeably.
[194] We adopt this terminology from Gibbard (2003), who takes normative judgments – the mental states that constitute one's answer to "what should I do?" – to themselves consist in intentions. Gibbard thus counts as a "rationalist" in our sense because he doesn't take states of the will to make any independent contribution to settling w.t.d., over and above that made by one's normative judgments. Indeed, Gibbard holds that it's simply a conceptual confusion to suppose that w.t.d. isn't settled by "what should I do?"

evaluative underdetermination. But these examples are unfortunate. The role of the will in these cases looks to be more of a hindrance (as in *akrasia*, where it seems to constitute "downstream" interference with deliberation's effect on behavior) or to be more of a mere randomization device (as in Buridan's-ass and other cases of evaluative/normative underdetermination).[195] Hence, even if the will does play an independent role in determining behavior, it looks to be a role outside the perspective that the (ideal) agent herself occupies in answering w.t.d., or, at best, to be a relatively peripheral capacity within it (rather than any central, important part of *her* qua agent).

To the extent that underdetermination has been discussed in decision theory, it's received similar treatment. Ullman-Margalit and Morgenbesser (1977), for instance, claim that Buridan's ass-like cases involve mere *picking*, rather than *choosing*, where only the latter type of selection is determined by one's preferences. In cases of picking, the agent herself may select an option, but only by being "transformed into a chance device that functions at random and effects arbitrary selections" (773).

Decision theory hasn't focused on the same questions about just what kinds of mental states constitute the sources of reasons-explanations that philosophy of action has, but we can translate those debates into its framework. The picking/choosing distinction makes clear that, in decision theory, answers to "what to do?" are constituted by one's *preferences*. The typical Humean view is that preferences are completely determined by (i) one's beliefs or credences, represented by a probability function, p, and (ii) one's desires (or the strengths thereof), represented by a utility function, u. The rationalist is most naturally interpreted as claiming that u represents evaluative/normative judgments (perhaps in addition to desires).[196] The volitionalist, in contrast, claims that there needs to be another, third component, in addition – one that represents the agent's will.

Translation in hand, we note that *recent work in decision theory provides direct, independently motivated support for volitionalism*. Specifically, *Risk-Weighted Expected Utility (REU) Theory* holds that, in cases of uncertainty, in addition to u and p, determining an agent's preferences also requires a risk function, $r(p)$ (Buchak, 2013). Buchak shows that $r(p)$ plays a genuine role in determining preferences, not just pickings. And, as we'll argue below, $r(p)$ seems to represent mental states of precisely the kind volitionalists tend to appeal to. Thus, REU theory shows that states of the will do make important, non-arbitrary contributions to ideally rational agents' answers to w.t.d?, over and above those made by their beliefs, desires, and evaluative judgments.

II. REU Theory

Traditional decision theory – expected utility (EU) theory – holds that p and u completely determine a rational agent's preference ordering: one must prefer the gamble with the highest average utility value, relative to p and u. By contrast, REU theory holds that it's (somewhat) up to the individual how to aggregate the utilities of the different outcomes she might receive in order to determine the value of the whole gamble: she need not average. Furthermore, two agents with the same utility and probability functions need not order gambles in the same way. In this sense, preferences and choices are underdetermined by one's beliefs and desires.

[195] In the classic example, Buridan's ass is equidistant from two indistinguishable bales of hay that it takes to be equally desirable, valuable, and that it takes there to be equal reasons to approach and eat. The ass supposedly starves, unable to make a choice, but many volitionalists have claimed that human beings in the same situation, at least, would not, and that this fact supports their position.

[196] We'll return to this claim below.

Consider a more complex case of underdetermination. Suppose you're asked to choose between two different sets of coin-flips. You can either take Deal 1, in which the first coin-flip will determine whether you get an Elvis stamp (heads, you win the stamp; tails, you get nothing) and the second whether you get a pair of gloves (tails, you win gloves; heads, nothing), or you can take Deal 2, in which the first coin-flip will determine both prizes (heads, you win the stamp; tails, gloves):

	HH	HT	TH	TT
Deal 1	Elvis stamp	Elvis stamp and gloves	Nothing	Gloves
Deal 2	Elvis stamp	Elvis stamp	Gloves	Gloves

The expected utility of both deals is identical (assuming the goods are independent, such that having one doesn't alter the value of having the other), so EU Theory requires you to be indifferent between the gambles. Buchak, however, shows that an ideally rational agent might prefer Deal 2 to Deal 1 because in the latter gamble, it's a sure thing she'll receive at least something – that is, because it's less *risky*. She might, in other words, care more about the worst-case scenario than the best-case scenario, even when these are equally likely and when the average utility of both deals is the same. Alternatively, she might prefer Deal 1 to Deal 2 because she cares more about the best-case scenario. Finally, she might, as EU theory recommends, be indifferent. Again, the point is that the ideally rational agent's choice is underdetermined by her utility and probability values.

EU theory requires choosing the gamble with the maximum expected utility. For a gamble $g = \{E_1, x_1; E_2, x_2; \ldots; E_n, x_n\}$,[197]

$$EU(g) = \sum_{i=1}^{n} p(E_i)u(x_i)$$

or, equivalently, where $u(x_1) \leq \ldots \leq u(x_n)$:

$$EU(g) = \sum_{j=1}^{n}\left(\sum_{i=j}^{n} p(E_i)\right)\left(u(x_j) - u(x_{j-1})\right)$$

The latter equation ranks the utilities of outcomes and then weights the differences between adjacent outcomes by the probability that an outcome at least as good as the better will obtain (assigning a weight of 1 to the worst outcome). That is, the value of a gamble is: its worst possible value; plus the interval difference between the worst value and the second worst value, weighted by the probability of getting at least the second worst value; plus the interval difference between the second worst and the third worst value, weighted by the probability of getting at least the third worst value; and so forth.

In contrast, REU theory holds that the weight given to these utility differences doesn't have to be identical to their probabilities, but is instead (somewhat) up to each individual. Thus,

[197] Following Savage's (1954/1972) framework, x_i is the outcome yielded by each (mutually exclusive and exhaustive) event E_i that might result from act *g*.

REU theory says to choose the gamble with the maximum risk-weighted expected utility. Where $u(x_1) \leq \ldots \leq u(x_n)$,

$$REU(g) = \sum_{j=1}^{n} r\left(\sum_{i=j}^{n} p(E_i)\right)\left(u(x_j) - u(x_{j-1})\right)$$

This equation ranks the utilities of outcomes and then weights the differences between adjacent outcomes by a *function* of the probability that an outcome at least as good as the better outcome will obtain. If this function, *r(p)*, is high, then the value of the better outcome will count for more (one will be risk-seeking); if low, the value of the worse outcome will count for more (one will be risk-averse) (Buchak 2013: 50). Thus, unlike in EU theory, on REU theory, (differences in) the utilities of outcome(s) are *weighted* differently in determining one's preferences depending on their relative rank.

EU and REU theory are both accounts of how the utilities of outcomes get aggregated to determine the overall utility of a gamble – to determine one's answer to w.t.d. EU theory claims that there's only one rational way to aggregate (EU is the special case of REU when $r(p) = p$). In this way, EU theory only allows agents to be sensitive to *local properties* of gambles – what happens in each outcome and the probability of that outcome obtaining. REU theory claims that rational agents can also be sensitive to *global properties* – e.g., to the distribution of utilities over a gamble's probability space, the extent to which the outcomes are spread out, its minimum, and its maximum.

By allowing sensitivity to global, and not just local properties, REU theory allows ideally rational agents to care about things like the fact that the worst outcome in Deal 1 – TH – is significantly worse than the worst outcome in Deal 2, and so allows them to prefer the latter. Indeed, the choice between the two deals above shows that to answer the question of w.t.d? – to determine one's preferences – it's often not enough to specify one's beliefs and desires. In addition, in cases of uncertainty, the ideally rational agent will also have to determine how to structure her aims, and will have to specify how to weight or balance the interests of her possible future selves (e.g., whether to ensure that the worst-off self at least gets something, or whether to ensure that her best-off possible future self is even better off – e.g., gets the stamp and the gloves – but at the risk of ending up with nothing). What *r(p)* represents is this balance (Buchak 2013: 55).

III. Volitionalism

Buchak holds that the risk-function represents "risk attitudes," which are not beliefs, desires or (we can add, and will argue in the next section) evaluative or normative judgments. The utility function already represents whatever attitudes one has toward the *values* of outcomes. And since outcomes are individuated finely enough to capture everything an agent cares about, it already represents whatever attitudes an agent has toward *reasons*, too. As discussed below, Buchak argues that one cannot re-characterize risk attitudes by taking global properties to be valuable in themselves or by taking them to alter the value of outcomes, and this is true regardless of whether the "taking to be valuable" is a matter of having desires or judgments (about values or reasons) – i.e., regardless of which attitudes we take to determine the question of "what should I do?"[198]

[198] See Buchak (2013: 36-47, 114-147).

One's risk attitudes determine how the values of outcomes contribute to a gamble. They don't represent values or reasons, but instead *weight* how much influence the values or reasons that there are get in determining one's preferences (that is, in determining one's answer to w.t.d?). Buchak notes that risk attitudes correspond to virtues recognized by folk psychology: being prudent/risk-averse (prioritizing a high minimum) or venturesome/risk-seeking (prioritizing a high maximum). These are arguably attitudes that folk psychology already recognizes to be states of the will, but they also seem to be precisely the sort of attitude that volitionalists in philosophy of action have focused on.

In particular, volitionalists tend to posit higher-order attitudes that have a "world-to-mind" direction of fit, but where the relevant parts of the "world" *are one's other mental states* – e.g., a higher-order desire for a first-order desire to move one to action (Frankfurt 1988), or an intention or *self-governing policy* to treat a first-order desire or consideration, *C*, as having a particular weight, *W*, in practical deliberation (Bratman 2007: 240). It seems clear that risk-attitudes are precisely this type of weighting of first-order considerations. They are represented by a function, *r(p)*, that takes probabilities of outcomes as inputs and, as output, weights the differences between the utility values of outcomes differently. In other words, *r(p)* represents mental states that weight how much one's desires (or other reasons-responsive states) count in practical deliberation – i.e., in determining w.t.d. In contrast, the attitudes that rationalists appeal to are first-order – i.e., responses to the properties (values or reasons) of outcomes themselves.[199]

Thus, REU theory seems to provide direct support for volitionalism, and support that doesn't suffer the shortcomings of traditional arguments. Bratman (2007), for instance, claims that we need self-governing policies to resolve impasses in potentially life-changing cases of underdetermination, like deciding whether to join the Free French or to care for one's ailing mother (Sartre 1956/1975). But worries about Buridan's-ass-like cases remain, such as their being arbitrary or relatively rare. Rationalists can try to claim that the self-governing policies in these cases are external to the *agent's* deliberative perspective, or that they still involve "mere picking" (even if picking of a potentially life-changing sort). However, REU theory shows that risk attitudes are required for resolving more complex cases of underdetermination (such as the coin-flip example) that don't involve "mere picking." Moreover, risk attitudes are required for resolving *nearly every* case of choice under uncertainty (the one exception being when one gamble "dominates," i.e., does better than every other regardless of which state of the world obtains). Being a rational agent involves deciding how to trade off the fact that one gamble does better in some state of the world against the fact that another gamble does better in a different state, even when the gambles' expected utilities are equal. Deciding how to make these tradeoffs is a fundamental feature of determining one's *preferences*, and in this way, REU theory demonstrates that self-governing policies are genuine constituents of an agent's answer to w.t.d?[200]

Risk attitudes are neither handicaps nor arbitrary selection accessories. The traits of being prudent and venturesome are stable dispositions that folk psychology recognizes as important constituents of one's identity *qua* practical agent. Nor is the contribution of risk-attitudes infrequent or confined to choices between indistinguishable bales of hay. Instead, given the prevalence of choice under uncertainty where no one gamble dominates, in giving reasons-explanations of (ideally rational) agents' behavior, we'll have to cite their risk attitudes – volitional states independent of their desires and evaluative judgments – *all the time*.

[199] See, e.g., Watson (1975) and Wallace (2006: 192-7).

[200] REU theory demonstrates *at least* this much. We leave open the question of whether the traditional, Buridan's-ass-like cases of underdetermination also involve preferences, rather than mere picking.

The existence of *risk-weighted akrasia* supports this conclusion. Consider again the choice involving the gloves and the Elvis stamp. The two deals have equal expected utility – the agent's reasons and values are, on average, equally forceful on either side. Suppose that the agent settles on being risk-averse, though, and so prefers Deal 2 to Deal 1. However, suppose the temptation to get both prizes, her fear of regret, or whatever mechanism of weakness of will you prefer, causes the agent to tell the bookie that she wants *Deal 1*. Even though the agent still acts in a way that maximizes her expected utility, she apparently acts akratically – in a paradigmatically irrational way – because she acts against her risk attitudes. Thus, acting rationally doesn't seem to be just a matter of being motivated by one's desires or evaluative/normative judgments in conjunction with one's beliefs or credences. Rather, on pain of *akrasia* and irrationality, one must act in accordance with, and be motivated by, one's volitional self-governing policies.

IV. Objections

The rationalist seems to be left with three lines of reply. First, she might accept REU theory but claim that risk attitudes are actually evaluative/normative judgments, rather than volitional states. But as discussed above, judgments of the sort that rationalists typically appeal to are first-order in a way that risk-attitudes are not.[201]

More plausibly, the rationalist might claim that risk attitudes are themselves *valuable* – i.e., that their presence or satisfaction in an outcome contributes to its value – and so are properly taken into account by one's evaluative judgments. There are two ways of developing this reply. First, the rationalist might claim that we've *misidentified* risk attitudes. She might claim that risk attitudes are actually just basic propensities, like (gustatory) tastes, rather than things we have any volitional control over. Like one's tastes, one's risk-related propensities might then be taken into account by one's evaluative/normative judgments. Plausibly, enjoying risk, or being risky, gives one a reason to seek and take risks, just like a taste for ice cream gives one a reason, or makes it valuable, to seek ice cream. A third, and related line of reply, is to accept that there are risk attitudes that are not just basic propensities – risk attitudes of the kind that REU theory describes – but to maintain that these attitudes are themselves valuable or generate reasons. We consider the misidentification reply first.

We accept that there may be basic risk-related propensities – "risk attitudes," if you wish – that do provide reasons to act or contribute to the overall value of an outcome (construed broadly, to include *any* values that obtain in it, including any that obtain in virtue of an agent's relation to the external state of affairs in that outcome). Folk psychology recognizes such risk-related propensities. For example, regret is a negative feeling that has negative value (Buchak 2013: 120). This might lead the rationalist to try to re-describe our coin-flip example above as the choice between:

	HH	HT	TH	TT
Deal 1	Elvis stamp	Elvis stamp and gloves	**Regret**	Gloves
Deal 2	Elvis stamp	Elvis stamp	Gloves	Gloves

[201] Moreover, this proposal would require evaluative/normative judgments that are underdetermined by reasons and values, since they occur in the face of underdetermination (equality) of expected utility. The Humean might also attempt to use any of these three strategies, but we'll focus on the rationalist.

Folk psychology also recognizes being assured of a good thing – a pleasurable anticipation or lack of anxiety that has positive value (120). In that case:

	HH	HT	TH	TT
Deal 1	Elvis stamp	Elvis stamp and gloves	Nothing	Gloves
Deal 2	Elvis stamp **and surety**	Elvis stamp **and surety**	Gloves **and surety**	Gloves **and surety**

Regret and surety are (un)pleasant emotions. They're features of outcomes that agents take to provide (dis)value or reasons "even once the gamble has been 'resolved'—once the risky event occurs and one outcome has been selected" (121).

In neither of these "re-individuated" choices does a preference for Deal 2 violate EU theory (or present any threat to rationalism or Humeanism). Because the outcomes are actually different, and have different values, the EU maximizer can consistently prefer Deal 2 to Deal 1. The way one will feel in (and about) an outcome once it obtains can affect its value, and so ground rational preferences for outcomes that are – independent of one's feelings or any other psychological attitudes had about the external states of affairs in those outcomes – exactly alike.[202]

We grant all this and accept that there may be basic risk-related propensities. But we maintain that risk-related attitudes properly so-called (i.e., of the sort described by REU theory) also exist. What Buchak (2013: Ch. 4) shows is that regret, surety, and other propensities that are (dis)valuable in themselves are not the only psychological responses that we have toward risk, nor the only such mental states recognized by folk psychology. In some cases, rational preferences between outcomes cannot be accounted for by using any such strategy of "local re-individuation."

For instance, we can stipulate that in the coin-flip example, your memory will be erased directly after choosing between deals (Buchak 2013: 132-3). Hence, in every possible outcome, you won't remember that you made any choice, and so can't regret that you didn't make it differently (or pleasurably anticipate the sure thing you're about to receive). You don't know that you made any decision at all – you just find yourself in the relevant state of affairs receiving the prize. In that case, no preference can be based on risk-related propensities like regret and surety, since the case now stipulates that the outcomes don't include these. Nonetheless, one can still rationally prefer Deal 2 to Deal 1. So, preferences still reveal the existence of risk attitudes, properly so-called. If you can prefer Deal 2 to Deal 1 even when the outcomes are stipulated to be identical – including in every psychological feature – then risk attitudes are required.[203] Even if there are risk-related propensities that figure into desires or judgments about outcomes, there are also risk attitudes that structure the weighting of one's desires in the way REU theory describes. And these are what the volitionalist has in mind.

---

[202] To make this strategy work in many cases, like the Allais paradox, though, the (dis)value of regret (or surety) must be made unrealistically extreme (Buchak 2013: 127-8).

[203] Even once the values and probabilities of all outcomes are known, there's still a further question about how to aggregate those features to determine a single value for the gamble (Buchak 2013: Ch. 4). Thus, even if there were no actual people with the preferences in question (even if everyone actually aggregated by setting $r(p) = p$ and so was "globally neutral") this wouldn't obviate the fact that there is still a component of decision-making beyond desires for (or judgments about) outcomes.

Especially once we distinguish them from risk-related propensities, it's exceedingly plausible to construe risk attitudes as volitional states – things that we can choose (or at least directly affect through choice). People may have brute feelings about regret and surety, but folk psychology also recognizes "calculated risk." When confronted with more complex cases of evaluative underdetermination, like the (memory-loss) coin-flip example above, one settles w.t.d. through first-person practical deliberation in a way that goes beyond any answer to the question of what values there are or what reasons one has. It's phenomenologically implausible to construe the attitudes that conclude these deliberations as beliefs or desires about what propensities one has, or as passively determining one's behavior. Rather, the agent *herself* calculatedly settles the question of "what to do" in these cases *by choosing* which risk attitudes to adopt.[204]

The third reply grants the existence of risk attitudes, properly so-called, but claims that these very attitudes give us reason to prefer certain outcomes. For instance, it might be that an outcome in which I act in accordance with my risk attitudes is a better, more valuable outcome, than one in which I don't. In other words, (i) planning to weight the value of outcomes in a certain way might itself be valuable, and (ii) the property of resulting from such a plan might contribute to the total value of an outcome. The latter need not be based on any positive or negative psychological feelings that obtain in the outcomes. Instead, they may be values or reasons that derive from the outcomes' histories, in which case this objection isn't answered by stipulating any memory loss.

However, the REU theorist and the volitionalist need not deny this reply, either. Perhaps having certain risk attitudes – *once* one has them – can provide reasons to act in accordance with those attitudes (rather than against them, or in accordance with other risk functions). But that's not the primary way in which risk attitudes figure into the determination of one's preferences. The primary role that risk attitudes play is in weighting one's first-order desires (or other reasons-responsive states), even if such weightings *then* give rise to differences in the values of outcomes (in which one does or does not act on them). An agent already has to weight outcomes in a given way in order for that weighting to give that agent reasons to prefer some outcomes to others.

Perhaps having certain risk attitudes – e.g., setting $r(p) = p^2$ gives one reason to act in ways that maximize REU according to that value, rather than others. And perhaps *once* one sets $r(p) = p^2$, then one has reason to continue setting it to p^2. Having those particular risk attitudes or acting on their basis might be more valuable than having, or acting on the basis of, the rejected alternative attitudes (e.g., of setting $r(p)$ to $p^{1/2}$).

But prior to choosing which ones to have, most risk attitudes will be *no more valuable* than any other. There's no more reason to set $r(p)$ to p^2 (or to act in ways that accord with that setting) rather than to set it to a different value – e.g., of p or $p^{1/2}$ – *before* one has already done so (viz., set it to p^2). One might be able to foresee that one will have most reason to act in ways that accord with $r(p) = p^2$ *if* one adopts that risk attitude. But if so, then one equally foresees having most reason to act in ways that accord with $r(p) = p$, or $p^{1/2}$, etc., *if* one chooses to adopt those weightings instead. Thus, even if one can anticipate that the way one weights outcomes will come to affect one's reasons or values, one *first* needs to choose to weight them *in that way* before the value or reasons of so weighting them can factor into one's deliberation in any way that can determine preferences. The initial choice of how to weight outcomes is still

[204] Alternatively, the agent might simply choose an action that he knows will maximize REU according to one risk function, $r(p)$, rather than others. In that case, he wouldn't directly choose between different risk attitudes, but (would understand that) his choice would determine which risk attitudes he thereby adopts.

evaluatively/normatively underdetermined, *even when* the values of having or acting on one's risk attitudes are included, since prior to settling on which to have, all of the relevant risk attitudes (within a certain range, at least) are equally valuable.

V. Conclusion

The volitionalist claims that more goes into motivating a rational agent than beliefs and desires – in particular, her second-order plans or intentions play an independent role in determining what to do. We have seen that risk attitudes fit this characterization exactly. In the first instance, the contribution they make to determining one's preferences and motivations is in a *volitional* capacity – *qua* plans or weightings of first-order states (even if they make a secondary, subsequent contribution to the values and reasons those states respond to). REU theory shows that the role of risk attitudes – and thus volitional states – extends far beyond Buridan's-ass-like cases of tie-breaking. *Whenever* one has to choose between gambles none of which dominates all others, in addition to fixing one's desires and all-things considered evaluative and normative judgments (i.e., settling "what should I do?"), one still has to fix one's risk attitudes (i.e., determine "how should I aggregate?") in order to answer "what to do?"

References

Bratman, M. 2007. *Structures of Agency*. Oxford: Oxford University Press.

Buchak, L. 2013. *Risk and Rationality*. Oxford: Oxford University Press.

Frankfurt, H. 1988. *The Importance of What We Care About.* New York: Cambridge University Press.

Gibbard, A. 2003. *Thinking How To Live.* Cambridge, MA: Harvard University Press.

Ullman-Margalit, E. and Morgenbesser, S. 1977. "Picking and Choosing." *Social Research* 44: 758-759.

Savage, L. 1954/1972. *The Foundations of Statistics.* New York: Dover.

Sartre, J. "Existentialism is a Humanism," in W. Kaufmann (ed.), *Existentialism from Dostoevsky to Sartre*. New York: Meridian/Penguin.

Wallace, R. J. 2006. *Normativity and the Will*. Oxford: Oxford University Press.

Watson, G. 1975. "Free Agency." *Journal of Philosophy* 72: 205-220.

Nicolas Cornell
University of Pennsylvania

Nicolas Cornell is an assistant professor in the Legal Studies and Business Ethics Department at the Wharton School, University of Pennsylvania. He has a J.D. magna cum laude from Harvard Law School, and he is currently in his final year of a Ph.D. in philosophy from Harvard University.

"Wrongs, Rights, and Remedial Ambiguity"
Commentator: Gina Schouten, Illinois State University

Abstract: *This paper argues that the nature of a wrong is not determined by the right that was violated. Many philosophers, such as Weinrib, Ripstein, and Darwall, maintain that rights and wrongs are necessarily tied to one another, reciprocal perspectives on the same moral connection between agents. This paper presents an argument for thinking that wrongs are not connected to rights in this straightforward way. The argument proceeds through an investigation of remedies. Remedies offer us a window into the nature of wrongs because they constitute what we consider the appropriate repair for the wrong. This paper shows that, in several ways, remedies do not correspond simply with the nature of the right that was violated. For any violation, other factors will figure in determining the appropriate remedy. In other words, wrongs—as viewed through the lens of what it would take to compensate for them—involve elements beyond rights.*

Introduction

Remedies aim to compensate the victim of a wrong. This is a foundational principle of private law, and it is also a moral principle. This idea connects wrongs with rights in an important way. We think that what it takes to remedy a wrong is a return of what was there ex ante—a restoration of the right that was taken away. Contemporary moral philosophers often view the connection between wronging and rights as illustrative of the relational character of morality.[205]

In this paper, I mean to challenge the straightforward understanding of what it means to remedy a wrong. I mean to show that the appropriate remedy for a wrong is not determined only by the nature of the right that was taken away. Rather, the appropriate remedy will depend on certain facts that are only present ex post—facts about what resulted and about how the transgression is interpreted. These additional facts are important because, from the ex post perspective in which wrongs arise, the idea of giving someone what was taken away is ambiguous.[206]

[205] See, e.g., Thompson (2004), Darwall (2009), Ripstein (2009), Weinrib (2012), Wallace (2013).
[206] The rough idea can be glimpsed in a quotation from Kaplow and Shavell (2009, p.166): "[T]he concept of compensation is ambiguous in the case of contracts: expectation damages make the victim of a breach whole by reference to a benchmark of performance, whereas reliance damages make the victim whole by reference to the position he would have occupied if not promise had been made." A promisee generally has a right that a promise be fulfilled. But the wrong of breaking a promise is not necessarily measured by how much the promisee expected to

I believe that this feature of remedies—that the appropriate remedy for a given wrong may depend on more than the nature of the right that was violated—shows that wrongs are qualitatively different from rights. That is, I am arguing that, even where there is a corresponding right, wrongs have a distinct character. Wrongs—as viewed through the lens of what it would take to compensate for them—involve factors beyond just the rights that were violated.

Blackstone wrote that, "[A]s all wrong may be considered as merely a privation of right, the one natural remedy for every species of wrong is being put in possession of that right" (1769, Bk.3, Ch.8). I mean to use this same reasoning in reverse—as *modus tollens* rather than *modus ponens*. If it is not true that every wrong has the single natural remedy of returning the right in question, then wrongs cannot be considered merely the privation of a right. That is, the complexity of remedies challenges the simple picture of wrongs as the mirror image of rights.

Remedies as Windows into Wrongs

A central premise of this reasoning—implied in both Blackstone's argument and in my own—is that the remedy for a wrong reflects the nature of that wrong. That is, the argument depends on the thought that remedies mirror the wrongs that they remedy. I begin by defending this premise, which motivates the focus on remedies in the remainder of the paper.

Remedies are remedies *for something*, and, as such, to know about a remedy is to know something about what is being remedied. A successful medical remedy tells us something about the nature of a physical injury that it cures. Similarly, a legal or moral remedy sheds light on the legal or moral injury that it cures.[207] And these remedies are illuminating because they correspond to the underlying injuries as an opposite, as a corrective.

Legal remedies have this same sort of strong correspondence with the injury being remedied.[208] Suppose that I have a legal grievance with my employer, and it is determined that the appropriate

---

gain. Sometimes, it will be better viewed in terms of injury resulting from having been given a bad promise to begin with. Which of these alternatives is appropriate will depend on context, and it may only be clear after the fact. So we cannot straightforwardly say that the remedy is dictated by the right that was violated.

[207] In this context, John Goldberg (2006) usefully distinguishes between two different meanings of "injury." Goldberg argues that the law originally understood injuries as wrongs but that an understanding of injuries as losses has arisen in modern times. The ambiguity that Goldberg describes might suggest that one cannot safely assume that remedies reflect wrongs. But, in fact, my methodology is generally consistent with Goldberg's argument. Goldberg demonstrates the shifting meanings of "injury" by, in part, demonstrating shifting judicial approaches to awarding remedies. That is, Goldberg's argument, like mine, assumes that looking at remedies tells us something about the conception of injury that is at work.

[208] This argument might seem to place too much stock in the concept of remedy. In doing so, it may seem to beg the question because we might wonder whether the legal ideas that we call ``remedies" are truly remedies in this sense. For example, Birks (2000) makes a plausible argument that the law would do well to replace talk of remedies with talk of secondary or remedial rights. But Birks's argument does not really challenge the idea that, whatever we call it, legal recourse reflects the problem being addressed. That is, Birks is not challenging the corrective function of private law, which is what my argument depends upon. In fact, Birks strongest argument is that the private law

remedy is six weeks of back pay. This remedy is revealing about the nature of the grievance. This illumination is based on the idea that legal remedies are corrective.[209] This idea is so intuitive that it seems almost built into the concept of a remedy, and, in the law, it seems built into the adage that remedies seek to make the victim whole again. If a different remedy had been appropriate, that would show that the grievance with my employer was different. That is, it would show that what was wrong was something different.

As long as this is true, we can use remedies as a window into the nature of wrongs. Put another way, thinking about remedies will be a way to think about the nature of wrongs. The remainder of this paper is devoted to using this strategy to think about the nature of wrongs. In particular, I mean to use the remedial question posed by the case of *Olwell v. Nye & Nisson*[210] to examine the conceptual composition of a wrong.

Restitution and Corrective Justice

Olwell v. Nye & Nisson

In 1940, Mr. E. L. Olwell sold his half interest in Puget Sound Egg Packers. As part of the agreement, Olwell retained full ownership of an egg-washing machine. The machine was stored in a storage space adjacent to the company. In 1941, the company had the machine removed from storage and put to use, without the knowledge or consent of Olwell. Upon discovering the unauthorized use four years later in 1945, Olwell offered to sell the machine to the company, but, after no agreement could be reached, he brought a legal action.

There is an array of conceivable ways that one might try to remedy the situation.[211] Olwell sought $25 per month, an amount aimed to recover the benefit that inured to the company as a result of the unauthorized use. The theory behind this legal action was that Olwell could "waive" his tort claim, and sue in "assumpsit" or quasi-contract instead. The trial court accepted

---

corrects things other than wrongs, and that talk of remedies can obscure this point because ``remedy" implies a wrong. Where private law is, in fact, addressing a wrong, Birks seems to have little intrinsic objection to the talk of remedies.

[209] This correspondence with the underlying injury is not to say that remedies are always complete. We often cannot fully remedy an injury. Physical injuries, for example, may be impossible to undo. In such cases, all we can offer is compensation, hoping to offer an equivalent level of well-being as a second-best alternative. The same, however, is true of medical remedies. A replacement joint, for example, may not fully restore its recipient. But it can provide the best functional equivalent available. In both cases, the fact that the recipient is never truly made the same as she was before does not mean that the response doesn't count as a remedy. They are remedies insofar as they are addressed to correcting the injury. They aspire to provide a restoration, even if they only partially succeed.

[210] 26 Wash.2d 282, 173 P.2d 652 (1946).

[211] Consider the following options: (1) return of the machine, (2) depreciation of the machine, (3) rental value of the machine, (4) the opportunity cost of being unable to use the machine, (5) restitution in the form of money the company saved through non-paid wages, and (6) restitution in the form of profits earned, (7) nominal damages in symbolic recognition of the violation. Any of these could plausibly be viewed as an appropriate way to respond to the company's nonconsensual use.

the remedy of disgorgement, and issued an award for $10 per week that the machine was used. This amount was calculated based on the wages for hand-washing that were avoided by using the machine.

On appeal, the company contended that Olwell had "an adequate remedy in an action at law for replevin or claim and delivery." In other words, the company contended that the appropriate remedy would be for them to give the machine back to Olwell. As such, the company argued that a suit in quasi-contract was inappropriate. Alternatively, the defendant company argued that if any damages were to be awarded, they should be the rental value of the machine, rather than a disgorgement of benefits. According to this argument, the appropriate remedy would be the amount that Olwell would have received from the use of the machine.

The Supreme Court of Washington rejected these arguments and affirmed the order for disgorgement. It focused on the fact that the plaintiff had the right to chose his claim. The logic of the decision, therefore, was clear: (1) Plaintiff "had an election"; (2) "Having so elected, he is entitled to the measure of restoration which accompanies the remedy." The Supreme Court did, however, modify the trial court's award insofar as it exceeded the amount requested. Olwell was awarded the $25 per month for which he had asked, not the $10 per week found by the trial court.

Weinrib and Corrective Justice

According Ernest Weinrib, awarding the company's profits did not conform with corrective justice. In Weinrib's view, corrective justice required that "the remedy reflect the wrong and that the wrong consist in a breach of duty by the defendant with respect to the plaintiff's right" (2001, p.20). That is, the remedy is taken to be a direct reflection of the legal right—the two are flip sides of the same coin. Any particular legal right creates an entitlement of one party with regard to another party. When that right is violated, the legal remedy is to give to the wronged party that to which they were entitled. Remedies are not ad hoc social instruments, but rather are part of the conception of a rights relation between two parties. Thus, the specific remedy must be "the notional equivalent at the remedial stage of the right that has been wrongly infringed" (2001, pp.4-5).

Based on his corrective conception of private law, Weinrib argues that the *Olwell* remedy was conceptually erroneous. Olwell's legal entitlement was to the machine. By using the machine without authorization, the company violated his entitlement to the exclusive use of his machine. The remedy should reflect the entitlement that was violated. As a result, Weinrib argues, the appropriate remedy is the fair market value of using the machine, i.e., the rental value.

By issuing disgorgement of benefits, Weinrib continues, the court assumes an improper framework of what it would mean to make the plaintiff whole. For Weinrib, the baseline comparison that was used was entirely confused: "Basing the damages in *Olwell* on the cost of hand-washing the eggs implies that the defendant was under an obligation to the plaintiff to wash the eggs by hand. This is absurd." (2001, p.20). For Weinrib, the disgorgement remedy would

suggest that Olwell had a right to the efficiency of using a machine over manual labor. According to Weinrib, by awarding profits, the *Olwell* court has stepped outside the bounds of private law. If it were correcting the rights violation, then it would have awarded the value of what was taken from the plaintiff by the defendant.

Ripstein and the Wrong of Use

Arthur Ripstein has not, as far as I know, written anything about the *Olwell* case itself. But he has discussed, from a Kantian perspective, how we ought to think about restitution damages. In *Force and Freedom*, he writes:

[S]ometimes a wrong will be completed, and if it is, its *effects* must be hindered in order to maintain the external freedom of the aggrieved party… [I]f I manage to enlist you in support of my projects without your consent, I must surrender to you any gains I make as a result. I must do so because your right to set your own ends must be treated as an embodiment of your freedom, and so given back to you… Using another's person or property without his or her permission is never consistent with freedom for all. Because the property exists for the benefit of its owner, the only way to redress another's use of it is to treat that use as though it were done solely for that person's benefit. (2009, pp.82-83)

In other words, the only way, in Ripstein's view, to remedy an unauthorized use is to give every benefit received from that use to the owner, as though the use were performed for his or her sake.

The basis for Ripstein's argument is that impeding an infringement of rights is itself a way to protect the freedom that rights safeguard. For this reason, rights are associated with an authorization to coerce. Coercion is authorized in such cases because "it restricts a restriction on freedom" (2009, p.55). He believes that this same idea of impeding a restriction on freedom explains remedial action as well. He writes, "The idea of the hindrance of a hindrance has a second, retrospective aspect to it as well. What is hindered in this case is not wrongful action but its impact on the external freedom of others." (2009, p.82). So the remedial action, when a wrong has been committed, is focused on removing the external impact of the wrong.

Where someone has used property without authorization, that person has appropriated the object to serve his or her own purposes. Ripstein believes that the way to remove the impact of this wrong is by treating the use as advancing the purposes of the owner. Whatever is acquired by unauthorized use must go back to the owner. This is the retrospective response that most hinders the hindrance placed on the owner's freedom. And what this means, in practice, is that the user must disgorge the gains obtained by the unauthorized use.

I think that the contrast between Weinrib and Ripstein is quite striking. It looks like Ripstein is saying that justice *requires* the remedy that Weinrib is calling *conceptual error*. This contrast is especially stark because Weinrib and Ripstein share many of the same commitments—in

particular, commitments to Kantianism and to corrective justice. What are we to make of their very different conclusions?[212]

Remedies as Essentially Ex Post

I mean to argue that the appropriate remedy depends not only on the nature of the right that was violated, but also on ex post features of the complaint and its context.

Dependence on Consequences for Degrees

The first point is rather simple. Whatever the conceptual basis for the remedy, the actual remedy will still depend on how much damage is done within that framework. If Weinrib is correct, then the *Olwell* remedy ought to have been the rental value of the machine. This amount could be more or less, depending on facts about the rental market. If, on the other hand, disgorgment of profits is correct, then the remedy will depend on the amount of profits that was realized. In short, whatever the framework, one must still determine the extent of the remedy required.

These inquiries mark a clear qualitative difference in the structure of rights violations and wrongs. Wrongs come in degrees in a way that rights violations do not. We can ask *how badly* was someone wronged? We want to know the *extent* or the *magnitude* of the wrong. We naturally speak about one wrong being *greater or lesser* than another. And this is true even where the rights violation is held constant. For any given right, a violation either occurs or does not occur. But the resulting wrong is not binary in this way; its can come in different degrees. If one believes that wrongs are conceptually equivalent to right violations, then such calculations and comparisons should be a puzzle. If the two are conceptual analogs, then why does one come in degrees in a way that the other not?

One might try to deny this disanalogy. It might be argued, for example, that the degrees in wrongs may be based on the importance of the right that was violated. While it is very plausible that wrongs can be greater or lesser based on the type of right that was violated, I find it

[212] There is some ambiguity in Ripstein's position that might reveal that the difference is less than it would appear. As I understand him, Ripstein is advocating for a disgorgement of profits. For him, whatever is done with unauthorized property should be treated as being done on the owner's behalf. So my sense is that, in the *Olwell* case, Ripstein would have awarded the profits that the company made by selling the eggs washed with Olwell's machine. This is what Weinrib rejects. But Ripstein might say that Olwell should receive only those profits attributable to the use of his machine and that anything else would overcompensate him. This, I suppose, would mean awarding Olwell whatever value was realized from the fact that the eggs were washed rather than unwashed. This would treat the washing of the eggs with the machine as though it were done solely for Olwell's benefit, and it would deprive the company of any gains from using the machine. This amount might be hard to calculate in practice if there is not a robust market for unwashed eggs, but no matter.

What is important for the present purposes is that, however it is interpreted, Ripstein's remedy would be conceptually quite different from two alternatives. First, the value contrasts with Weinrib's suggestion that Olwell receive the rental value of the machine. And second, it contrasts with disgorging the benefit received by the company, namely not having to pay for hand-washing the eggs. In an efficient market, all of these values might converge on one another, but this is essentially irrelevant. For one thing, we need to know what remedy to award in the real world, where markets are not perfectly efficient. More importantly, this hypothetical convergence does little to alleviate the conceptual difference between Weinrib and Ripstein. Even if they could arrive at the same dollar amount, their rationales would be discordant. This discord is, as I have said, especially noteworthy because one would think that Weinrib and Ripstein's views should be quite harmonious.

implausible to say that this can account for all differences in degrees that wrongs come in. Even when the exact same right is violated, the wrong will be greater if greater damage has been inflicted. This is the familiar phenomenon of resultant moral luck.[213] One person might appropriate a machine that would only have sat in a warehouse otherwise; another person might appropriate a machine that, as things turn out, became desperately needed by its owner. The latter person has violated the same right, and yet he has committed a greater wrong, as reflected in the greater remedy that is owed.

Taking the opposite tack, it might be argued that, in fact, wrongs are binary like rights violations. On such a view, it is injuries—the harms that result from wrongs—that come in degrees, not the wrongs themselves. The trouble with this reaction is that it disconnects wrongs from remedies—and from our other moral practices and experiences. Wrongs are simply stipulated to be the placeholders where rights are violated. We, of course, can use the term in this way if we so choose, but we are still left with something unaccounted for. There is *something else* that is what we are trying to remedy (and what we blame, resent, apologize for, forgive, and so on).[214] My interest is this something, whether we call it the wrong or not.

Dependence on Consequences for the Form of the Remedy

The previous section argued that there is a qualitative difference between violations and wrongs insofar as wrongs come in degrees, which are dictated in part by the consequences ex post. That argument, however, is compatible with the corrective justice view that the nature of rights dictate the form that remedies should take. In this section, I mean to call this view into question as well.

As witnessed already, there is not always consensus about what remedy is appropriate as a matter of corrective justice. Weinrib and Ripstein and the *Olwell* court all offer contradictory views about the appropriate remedy in the *Olwell* case. The reader may have an intuitive opinion about which of these is correct.

Regardless of which reaction one has to the *Olwell* case, I want to suggest that that reaction will not be consistent across all cases. Sometimes the rental value is too weak, and sometimes disgorging profits is too strong. If this is true, then conceptual analysis of the right will not

[213]As Nagel describes it: "If someone has had too much to drink and his car swerves on to the sidewalk, he can count himself morally lucky if there are no pedestrians in its path. If there were, he would be to blame for their deaths, and would probably be prosecuted for manslaughter. But if he hurts no one, although his recklessness is exactly the same, he is guilty of a far less series legal offense and will certainly reproach himself and be reproached by others much less severely." (1979, p.29).

[214]This something else can't just be injury in the purely descriptive sense of harm or loss. Many losses or harms are not actionable at law and are not wrongs morally speaking either. One might say that we are remedying wrongful losses. But this only begs the question. As Weinrib (2012, pp.121-23) points out, this phrase can be understood in two different senses. If one means losses that resulted from wrongful acts, then the moral significance of this category must be explained. If one means wrongful loss in Weinrib's normative sense, then it starts to look like the rights violation.

provide us with an appropriate remedy a priori. Wrongs include an extra element, not present in the right itself.

Consider two ends of the spectrum. First, imagine that, instead of a piece of heavy-duty, labor-saving equipment, the wrongfully used item had been merely a pencil, used to write out the business plan. No one would seriously think that, because of this wrongful usage, Olwell should be entitled to all the profits that resulted from that business plan.

At the other end of the spectrum, imagine that Olwell had owned a magical goose that would very occasionally lay golden eggs when it was caressed.[215] Imagine that the goose only produced eggs rarely and at random, making the expected value of the goose on any given day quite low. If Olwell's magic goose were taken from him for an hour, during which time it laid a large golden egg for the thief, it would be odd to say that he is only entitled to an hour's rental value of his goose.

What varies in these examples is the relative contribution of the appropriated item and the appropriator's efforts. This contrast might suggest a principle: each party should receive the equivalent of what he or she put in. As appealing as it sounds, this suggestion ignores the fact that the inputs combine to create something new. We cannot reverse time, and we cannot say precisely the effect that various forces had in getting us to where we are.

Consider the *Olwell* case. The company's labor combines with Olwell's machine to produce revenue. Insofar as the venture was worthwhile, the revenues will be more than sum of the value of the inputs. So we cannot merely give the value of the machine's use back to Olwell and the value of the labor back to the company, because there will still be more left over—what was *created* by the productive activity.

We might think that the surplus can also be distributed based on the relative inputs, either proportionately or entirely to the larger contributor. Thus the pencil owner would get little or none, whereas the goose owner would get most or all. But even this principle is inadequate. Suppose that the goose was stolen, but it did not lay an egg. Should the owner receive any compensation if he sues? Most people would think that he should. His rights were violated, and he deserves something representing the fact that he was dispossessed of his property. He should not be denied any remedy just because the thief didn't profit from his crime.

If the goose owner should get the benefits if there are some but should get a different remedy if there was no benefit produced, then it seems that the appropriate remedy for the goose owner cannot be determined ex ante. It depends on what the appropriation of the goose yielded. If it was fruitful, then the owner is entitled to those fruits. If not, then the owner is entitled to the

[215] If this example is too fanciful, the reader can substitute a lottery ticket or a copyright in the argument that follows. What is important is that the property's value is uncertain ex ante such that its actual value when put to use may significantly exceed the rental value.

rental value. One might think something similar about the *Olwell* case as well: If the company made a fortune by stealing his machine, perhaps he should get a share of that fortune. But if the venture was a complete failure, that shouldn't prevent him from getting any compensation.

It might look like there is still an ex ante principle here: the goose owner gets either the profits or the rental value, whichever is greater. Although we might say this ex ante, it is not a principle that determines the remedy based only on what is present ex ante. Even if we fully understand the rights involved, we do not know ex ante what will appropriately compensate a violation of that right. In this sense, the content of the wrong depends on something other than the content of the right.

Dependence on How the Grievance is Framed

Above, I attempted to show that the appropriate remedy is not given ex ante because it depends on what results. In some contexts at least, the plaintiff would seem to be owed either the profits or the rental value of the property, whichever turns out to be greater. But even this description, I believe, ignores an important way in which the appropriate remedy depends on something ex post. In this section, I mean to point out another way in which the remedial question is dependent on the ex post context: it depends on how the plaintiff frames his or her complaint.

Normally, if one has a choice between two amounts of money, one will choose the greater sum. So, where a plaintiff can seek either profits or rental value, we can normally expect the plaintiff to seek the greater amount. But the appropriate remedy isn't actually dependent on which amount turns out to be greater. Rather, it depends on what remedy the plaintiff requests, which in turn depends on how the plaintiff frames his or her complaint.

This point is clearest when the possible remedies are not both fungible, monetary values. Imagine that an employer makes an employee work extra hours on a project outside the scope of the employment contract. The employee comes up with an innovative idea, which the employer promptly patents.

We can imagine the employee seeking either compensation for his uncompensated hours of labor or seeking ownership of the patent. These represent different forms of complaint against the employer. One is a complaint that one didn't get paid; the other is a complaint that one's idea was stolen. The same set of facts could be basis for either grievance. It depends on how the employee perceives or frames the wrong.

In this sense, the appropriate remedy depends on the choice of the employee, which need not correspond with greater economic value. The employee might elect to seek lost pay rather than the patent itself, even if that is worth more. Perhaps he does not want to jeopardize his employment relationship with the company. Or perhaps he is not interested in having to license and police the patent himself. On the other hand, an employee might seek the patent, even where its market value was less than the wages would amount to, if he was particularly attached to its being *his idea*.

What these possibilities show is that a plaintiff doesn't simply receive (if successful) whichever remedy is greater. He receives the remedy that he elects to seek. Put another way, the appropriate remedy depends on how the complaining party frames the complaint. The remedy, that is, depends in part on how the victim perceives the injury. This is necessarily determined ex post. We may be able to speculate ex ante about how some action is likely to be regarded—what the complaint will probably look like—but it is the way that it is actually regarded—what the complaint actually is—that matters. This is a third way in which the remedial question depends on something ex post, not simply on the nature of the right ex ante.

Corrective justice theorists—Weinrib chief among them—rightly criticize non-corrective accounts of tort law for not capturing the character of private law that involves doing justice between the parties.[216] They claim that one cannot understand the privateness of private law without appreciating the bipolarity of corrective justice. I think that there is much to be said for this criticism.

But it is not true that non-corrective views have no conception of private law. As one writer puts the point, "[p]rivate law is structured as a drama between plaintiff and defendant" (Dagan 1999, p.147). What makes private law private is the structural fact that it adjudicates complaints of one party against another party. That is, the distinctive character of private law comes from the structure of relying on private complaints. In private lawsuit, one party makes a complaint against another. This is, in a sense, an assertion that "you have done me wrong." The defendant, then, is put is a position of responding to this complaint. What is important about this fact is that private law is not about resolving rights violations, but rather about resolving complaints. Private law does not attempt to correct for every rights violation. Only once a complaint has been raised does the law come to bear. Weinrib's account suggests that private law corrects moral imbalances. But this is not entirely correct—there must also be a complaint. Private law is about the relationship between the parties not just in its content, as Weinrib would have it, but also in its structure.

This is where Weinrib's account of the *Olwell* case goes awry. Weinrib suggests that the only question for the court concerned the nature of the right that was violated. But more immediately, the appropriate question concerned whether Olwell's complaint was a successful one. Olwell's complaint, put simply, was as follows: "You wrongfully stole the benefits of my machine—you owe them back to me." As the court notes, he might have made a different complaint. For example, he might have said essentially, "You stole the use my machine—you owe me the cost

[216]For example, Weinrib writes, "Presenting corrective justice as a quantitative equality captures the basic feature of private law: a particular plaintiff sues a particular defendant. Unjust gain and loss are not mutually independent changes in the parties' holdings; if they were, the loss and the gain could be restores by two independent operations. But because the plaintiff has lost what the defendant has gained, a single liability links the particular person who gained to the particular person who lost." (1995, p.63).

of using it," which would have been the complaint Weinrib imagines. But the former, not the latter, is the complaint that he elected to make in light of the facts available ex post.

The question, then, is whether the egg company can rebut the complaint that is made against it. Put simply, the question is whether the company can respond, "No, these benefits are rightly ours." This response wears its difficulty on its sleeve. The company cannot make this claim; from their mouth it is implausible. The court's opinion makes precisely this point: "However plausible, the appellant *cannot be heard to say* that his wrongful invasion of the respondent's property right to exclusive use is not a loss compensable in law."[217]

The *Olwell* decision is attentive to the complaint-based character of private law—to how the parties framed the dispute—in another way as well. Mr. Olwell could not be given more than he sought. Olwell's complaint was essentially "you owe me $900." It was therefore judged to be an error for the trial court to say, "He's right, you owe him $1,560." The court was limited by how the parties frame the dispute.

The structural dependence on how the parties litigate the dispute—that is, on how the grievance between the parties is framed and rebutted—is an important way in which the remedial question is not just transparent back to the right that was violated. It is based on that distinctive feature of private law as involving with one party addressing another party. These addresses are necessarily made ex post. And they demonstrate an important way in which remedies, and the wrongs that they remedy, are necessarily ex post.

Summing Up: The Unavoidable Ambiguity in Compensation

We typically think that compensation means giving back to a wronged party whatever was taken from him or her. We speak of making someone whole. In this picture, the wrong is the void that must be filled. It is as though a piece of a puzzle has been removed and just needs to be put back in. What to put back is the same thing that was there before it was removed.

But matters are not quite that simple. Things look different at one time than they do at another; events change things. Repairing a puzzle is not straightforward when the pieces can merge and morph. One is not just trying to return whatever was there before. Instead, repair must involve something new, which is based both on what was there before and on how things look now.

Wrongs have this character. They may arise from the fact that a right has been taken from us, but their shape is not determined only by the shape of the right that was taken. It also depends on the context ex post. That is, the nature of a wrong depends on certain facts that only come into existence once the wrong is committed. These include facts about the losses of the wronged party, the benefits derived by the wrongdoer, and the wronged party's interpretation of his or her injury. As a result, wrongs are not just the conceptual antipode to rights. Wrongs—as viewed through the remedies that they demand—are also a function of context and consequences.

[217] 26 Wash.2d at 286 (emphasis added).

References

Birks, P. (2000). "Rights, Wrongs, and Remedies." *Oxford Journal of Legal Studies*, 20(1):1–37.

Blackstone, W. (1769). *Commentaries on the Laws of England*. Clarendon Press.

Dagan, H. (1999). "The Distributive Foundation of Corrective Justice" *Michigan Law Review*, 98(1):138.

Darwall, S. (2009). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Harvard University Press.

Goldberg, J. C. (2006). "Two Conceptions of Tort Damages: Fair v. Full Compensation." *DePaul Law Review*, 55:435–468.

Kaplow, L. and Shavell, S. (2009). *Fairness versus Welfare*. Harvard University Press.

Nagel, T. (1979). "Moral Luck." In *Mortal Questions*. Cambridge University Press.

Ripstein, A. (2009). *Force and Freedom: Kant's Legal and Political Philosophy.* Harvard University Press, Cambridge, Mass.

——— (2011). "Civil Recourse and the Separation of Wrongs and Remedies." *Florida State Law Review* 39(1): 163-208.

Thompson, M. (2004). "What is it to Wrong Someone? A Puzzle about Justice." In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz.* Clarendon Press.

Wallace, R. J. (2013). "The Deontic Structure of Morality." In Bakhurst, D., Little, M. O., and Hooker, B., editors, *Thinking About Reasons: Themes from the Philosophy of Jonathan Dancy.* Oxford University Press, Oxford.

Weinrib, E. J. (1995). *The Idea of Private Law*. Oxford University Press.

——— (2001). "Restitutionary Damages as Corrective Justice." *Theoretical Inquiries in Law*, 1(1).

——— (2012). *Corrective Justice*. Oxford University Press.

Gary Watson
University of Southern California

Gary Watson is the Provost Professor of Philosophy and Law at the University of Southern California. His interests include agency, responsibility, and freedom of action and the will. A selection of his essays appeared in 2004 under the title *Agency and Answerability*.

"On The Importance of Having a Life?"

Commentator: Edward Hinchman, University of Wisconsin – Milwaukee

Intro…

I Having a Life

Thomas Reid writes at the beginning of Essay III of his *Essays On the Active Powers of the Human Mind* (1788):

> "It is a most important part of the philosophy of the human mind, to have a distinct and just view of the various principles of action, which the Author of our being has planted in our nature, to arrange them properly and to assign to every one its rank[218]….By this it is, that we may discover the end of our being, and the part which is assigned us on the theatre of life." (III.1.1, p. 95).

At one level of abstraction, all rational creatures have the same "assigned part", in the sense Reid has in mind: to realize and to guide our lives by the "best ends" in our circumstances for a whole life. That's the "end of our being". But the principles of self-regard and duty do not determine our individual "plans of life," so there is much that each of us must figure out on our own. For example, we must devise a general "system of conduct", and cultivate "habits of the will"(II.iii,88) that are well suited to realize the "best ends" within the circumstances and limitations of our particular lives. We are endowed with "reason and conscience" for this purpose, which enable us to apprehend the rational principles of self-regard and duty. Self-regard is "regard to our good [that is, happiness and perfection] on the whole; which must be estimated…from a due consideration of [our actions'] consequences…during the whole of our existence" (III.ii. 204). Together with the principle that pertains to just and benevolent conduct, our habits of will and system of conduct enable us to "regulate the inferior" motivational dispositions "so that they may conspire, in a regular and consistent plan of life, in pursuit of some worthy end." In this way rational creatures are capable of self-government, which consists in "lay[ing] down a rule to themselves which they are not to transgress, though prompted by appetite, or ruffled by passion". III.iii.8, 251.

Devising a suitable "system of ends" requires us to commit ourselves to vocations and to social and economic roles that structure our pursuits and sensibilities. To pick up the political metaphor of

[218] What is ranked here is not people but "principles of action." Principles of actions are for Reid sources of motives ("everything that incites us to act", 95). In III.i, Reid notes that "different principles of action have different degrees of dignity, and rise one above the other in our estimation, when we make them objects of contemplation." 159

'self-government', we must not only comply with the principles that are revealed to us as rational beings, we must ourselves legislate general rules (in the form of "resolutions") by which to organize a human life in the circumstances in which we find ourselves.

Reid thinks of the achievement of self-government, and especially the adoption and pursuit of general ends of life, as integral to *character* ; someone who has no fixed purposes has "no character at all" (88). Such a person will "be honest or dishonest, benevolent, or malicious, compassionate or cruel, as the tide of his passions and affections drives him." Without guiding "fixed purposes", "there would be no consistency in his conduct. He would be like a ship in the ocean, which is bound to no port, under no government, but left to the mercy of winds and tides." II.iii. 89. Continuing with this trope, Reid says that

> "it belongs to the rational part of our nature to intend a certain port, as the end of the voyage of life; to take the advantage of winds and tides when they are favourable, and to bear up against them when they are unfavourable." (III.v.222. See also III.viii. 198.)[219]

(In a letter from about the same time Reid wrote the *Essasys*, Mozart gave voice to a much more specific view of our (or at least his) station and its duties. "We live in this world to compel ourselves industriously to enlighten one another by means of reasoning to apply ourselves always to carrying forward the sciences and the arts.") (A letter to Padre Martini in 1776) [220]

It's a good question how much of Reid's picture depends upon his theism. Clearly much of it does. But even those of us who dwell in a "fatherless universe" (to use a phrase of Adam Smith's) find ourselves with lives to lead as well; many of the themes and questions Reid articulates are utterly familiar in non-theistic conceptions as well. Reid's picture involves the following claims, or something like them.

(1) That our being or existence has an end.

(2) that it is a normative imperative to live as good a life as we can

(3) that this requires thinking about our lives *on the whole*, both temporally and in its various dimensions.

(4) that this requires identifying and having "plans of life", or "systems of ends", conceptions of the good for one's life on the whole.

(5) that having a life is having a life *to lead*.

---

[219] Reid thinks that we are "in the image of God" insofar as we are self-governing. (IV.5, p.301). The point is not, absurdly, that God too manages his appetites, but that we author rules by which we order nature including our own impulses and relevant parts of our environment.

[220] From a letter to Padre Martini in 1776. Quoted in Wiggins, "Truth, Invention and the Meaning of Life", 89. Wiggins comments: "The foundation of what we envy [in the sentiments expressed here] was the now (I think) almost unattainable conviction that there exists a God whose purpose ordains certain specific duties for all men, and appoints particular men to particular roles or vocations."

(6) that leading a life consists not only in guiding our conduct by our "system of ends", but in some way fashioning that system as we go along.

(7) That to lack a guiding system of ends is to lack something (Reid calls it 'character') that we think of as distinguishing us from lower orders of animals.

I won't able to discuss all of these claims, or many of them in detail. But I want to focus on some of them in what follows, beginning with the idea of taking an interest in life as a whole.

ii Life as a whole

Many will agree with the second claim—they may find it axiomatic or even trivial. A putatively trivial version might be: it is best to lead as good a life as one can. Another version might be: it is one's duty to live as good a life as one can. Both of these need qualification: so long as one's life at its best is good enough, that is, worth living. This qualification of course broaches the fundamental question whether one should live at all, whether it is good for me to (continue) to have a life. The second form of the question is problematic if the duty or obligation is meant to be "directed", an obligation to others: this is one form of the question that arguably can't survive outside the theistic context like Reid's. In that context, one is held to account at the Last Assizes for how one's lived. Outside that context, it is doubtful that one has to account for oneself to anyone else for that question. Who else's business is it? But to some, it still makes sense to think that at least that *is* one's own business, that for example at the end of one's life self-rebuke is warranted to the extent that one has largely wasted one's time.

Is this thought a non-theistic version of Reid's idea of "the end of our being"(the first claim)? That end for Reid is something like entering the Kingdom of Heaven or union with god. Living as good a life as one can might be seen as instrumental to this, or even defined in terms of that. Here the 'end' of life means its purpose or *telos*.

I want to pursue this question a bit in terms of Reid's simile of life as a voyage or journey.

We don't set out on life's journey but rather find ourselves already underway. We can then ask the question, where do we want to go from here, or again whether we want to continue at all. If you thought this life possibly led to another, further life, it would make some sense for you to think of your life as a journey to that ultimate destination. In Reid's image, your final port, if you succeed, is a portal to paradise. Are there intelligible non-theistic versions of the idea of life as a journey? Think of life as a more or less lengthy trek. Your trek might aim for the highest summit, or, better, the highest number of high summits. You take that aim to be the measure of how well you lived. Like the theistic version, your life would then be governed by a dominant end. But I take it that most non-theistic lives have no such end. They are like treks with no overall destination. That is not of course to say that they are pointless, though some might be, and though some theists would make that objection of all such treks. The trek as a whole, which begins when you purposely hit the trail, and ends when you can walk no further, consists in shorter walks with many different destinations. There is no place to which all these shorter journeys are meant to lead, so in a way the trek as a whole is in an important way not itself in the end a single journey; it is simply constituted by the series of your treks. The map of where you've been and where you've ended up is a retrospective representation rather than a guide.

Possibly, you have a final destination in the innocuous sense of a place at which you wanted to end your travels—in the mist of an especially breathtaking waterfall whose spectacle you imagine as your last vista. This purpose may have (provisionally) fixed the terminal point on the map, and

influenced which other destinations you pursued (since you wanted to leave yourself in a position to reach this location), but it was after all just one purpose among others in the complicated course of your journey, not the "end" of your journeying except in a spatial and temporal sense.

Again, unlike you, some may have an end of their journeying as such, for instance they are "summit baggers". That may be the reason they're on the trail at all. Just so, some may think of their lives in this way; their existence has an end: to make as much money as possible, to serve the poor, to win the Nobel Prize—or union with God. Certain people, it seems, *live for* philosophy, or painting or family, meaning that's their life focus. These may be "ground projects" in Bertrand Williams's sense; without these projects one may see no reason to go on. I think describing some lives this way is very often hyperbolic. In any case failing to have dominant ends or a ground projects in this sense is not, I think, to the mark of deficient life.

Yet isn't there something more that we have, and think we ought to have, as a final end in life, even on a secular view—namely, as the second claim has it, the grand purpose of living the best life we can-- In terms of our simile, of putting together and executing the best overall package of treks one can put together (given the weather and one's personal limitations)? (As Ronald Dworkin put it, "We have the abstract ambition to lead a good life, and we worry, some of us all of our lives, about what that is."[221] It is presumably impossible to map it all out in advance (even a gifted and fully informed tour guide would have to deal with contingencies by forming back up plans), life is too contingent, and it is arguably undesirable to spend too much time thinking about this, because that is apt to spoil the quality of one's travels. Just the same, at the end of the journey, one wants to be able to say that one has travelled well overall. After all, the answer to what is conventionally taken to be the canonical question of ethics—Socrates's question, how we ought to live?—seems to involve a conception of a choiceworthy life overall, not just in some respects or for the near future. [Distinguish between "thinking about one's life in some respects and in the relatively near future is enough" and "one's life in some respects and in the relatively near future is all that matters."]

But it is mistake to think that an answer to this question is committed to the idea that over and above the choiceworthiness of one's ends, and relationships, and experiences, there is necessarily a further question about the choiceworthiness of the life one has led. The goodness of a life is just the goodness of the activities, relationships and experiences within it. (Unsurprisingly, this is just another way of stating the distinction invoked in the interpretation of Aristotle's idea of *eudaimonia* as an end: as an "inclusive" rather than "dominant" end, as a "formal" rather than "substantive" end. I take no stand here on the interpretative question.)

(ii) *On **not** having a life*

It will help to clarify the topic of the importance of having a life to ask what it might mean not to have one. How can one fail to have a life, except by dying or becoming irrecoverably comatose? Having a life in the sense that interests us not just the idea of a human being's psycho-physical span of existence, one's being alive as a human being. One can lose one's life, or fail to have a life, without ceasing to live.

Consider the life you had before getting a divorce, or going to prison, or contracting a severe illness, or becoming a refugee whose relocates to an alien culture (a Somalian in

[221] Ronald Dworkin, 205 *Life's Dominion.*

Minnesota) or having a religious conversion (or unconversion). You might strive to "get your life back" or despair of doing so, or you might see yourself as in a better place.

The loss of one's former life could be more or less global. One could lose one's domestic life or professional life, or social life, or love life, or life as a free citizen. In this usage, the life you have (at a period) refers to a constellation of relationships, commitments, plans, vocations which are central to how your agency and experience during some part of your existence is focused. In many of these cases, the loss of one's former life is replaced by a "new life", that is, a new set of engagements and attachments and pursuits. In this sense, it is common for the period of one's being alive, one's existence as a psycho-physical human being, to comprise several "lives" over time. Indeed, this is inevitable if only because of differences in developmental life-stages. If one lives long enough, one leaves one's life as a child behind, enters one's "prime" (though how life stages are conceived and lived is culturally diverse) and becomes an "elder" whose past is much longer than one's future, and whose future is limited not only in years but by alterations in one's vital powers. The lives available at these stages are bound to some extent to be different. (Again, the extent and nature of these differences is very culturally variable.)

These are examples of a single person having several lives diachronically, but we also speak of people "living two lives" during the same period. This needn't involve pathology, or the duplicity of the double-agent or the bigamist. For some of us, different synchronic parts of our lives are so disparate in the sensibilities that they require and express that it feels to the agents as though they occupy two different "worlds".

In sum, I've been emphasizing the distinction between having life and having a life. Any living, conscious reflective human being has life. But this is insufficient for having a life as a structure of commitments, engagements and relationships. The topic of this paper is the importance of having that. Reid is right to say that it is these that give character to the agent and provide the central content of someone's life as a human being. Biographies are largely about this content, I suppose, though that is a trickier question than might first appear.

What counts as "part of a person's life" is rather indeterminate. Whether or not an occurrence on the other side of the globe is part of her life depends on how it "impinges" on her, but that of course is a contestable notion. Does "impinging" require registering on her awareness, on her interests determined more largely? Is it part of her life that her son, from whom she is estranged but still cares about, dies in from a drug overdose in Prague even though she doesn't live to hear the news? In a literal biography, I suppose, that item would deserve mention, but perhaps not much, for it wouldn't affect how she goes on with things. And this might be a pretty good characterization of impingement: what affects how she goes on with things. But her son's death would dash her hopes for reconciliation, whether she knows it or not. But that is still too general. Some obscure climate change might affect her finances in a way that affects "how she goes on", or "how she responds to what happens (or she thinks happens)". Virtually anything in the causal order might be part of one's life in this sense. So the idea of "what belongs" to a life is really quite unruly. One thing is clear. To care about one's life as a whole is not an attitude toward something that is fixed entirely independently of one's concerns. In acquiring and cultivating cares and interests, one is shaping the contours of one's life, determining what counts as part of it.

(ii) *Depression, ground projects…*

In a clear sense, the lives of human beings have content, whether or not they meet the conditions for having a life. Consider Carlos who spends his mornings in dark melancholy under the bed covers. He ignores the door bell and the ringing phone. His family hasn't been able to reach him. He occasionally sits up to idly flip through the TV channels with his remote, weeping all the while. His brother climbs through a window in the upstairs bedroom, and prods Carlos to take his medicine or to come with him to the hospital. He hasn't worked for months, has lost his job, his wife feels abandoned and has moved to another town in order to support the child that Carlos used to love but now barely thinks about. Carlos's problem is not that his ground projects have been obstructed or ruined, but rather that he's ceased to have any.

Even though he doesn't have a life in the sense I've identified, Carlos's biography would surely include material of this sort. Of course it's important to understanding what's going on that Carlos had a life, to which there is hope for recovery (though the hope is not his). And in some sense the fact that his family cares about him, and that Carlos is impervious to that, is part of the content of his life.

Carlos is in a bad way. This of course has to do with his deep suffering. But it also has to do with the fact that he lost his life and is unable to get it back or to start anew. On a variation of the story, it may be that the life he had was bad for him. These examples seem to show that the importance of having a life is not just a function of the substance of that live. It may be bad for him that he lost the particular life that he had, but he needs to have a life of some kind. Even if we imagine that the medication prevented anguish, in so far as Carlos cares about nothing, he is in a bad way for a human being to be. He would be better off, by far, if he had a life. For that reason, those who care for him want this for him.

What these examples also show is that a person's being well or badly off is not just a function of what he cares about. I'm imagining that Carlos no longer cares about having or getting a life. He doesn't care about anything. That's a bad way to be. [Note on Frankfurt]

[The abusive sense of "get a life". Here the judgment may not be that the person fails to have a life but they fail to have what we might consider a meaningful life. (See e.g. Susan Wolf.)]

(iii) the estranged, psychopaths.

II The importance of *leading* a life

i) conceptions of the good

I turn now to some questions about leading a life and the connection of that with what Reid calls "systems of ends". For Reid, this connected with "self-government". For him this has to do with implementing a plan of life which requires controlling our lower motivational dispositions. But self-government is not just a matter of policing those dispositions, but of forming some of the commitments that are to be enforced.

It is natural to put Reid's notion of systems of ends in terms of the idea of "a conception of the good in something like John Rawls' sense:

> [A] conception of the good normally consists of a more or less determinate scheme of final ends, that is, ends we want to realize for their own sake, as well as of attachments to other persons and loyalties to various groups and associations. These attachments and loyalties give rise to affections and devotions, and therefore the flourishing of the persons and associations who are the objects of these sentiments is also part of our conception of the good.[222]

These notions play a similar role in these philosophers' practical philosophies. In Reid's case, the capacity to form and implement a system of ends demarcates moral agents from the lower orders of animals. For Rawls', having a conception of the good (or "what is valuable in human life") is one of the two moral powers distinctive of persons.

Charitably understood, conceptions of the good are instantiated multifariously, not only in judgment and deliberation but in attachments, affections, trained appetites, and habits of will which constitute a kind of "second nature." Just the same, they are implicitly second-order, inasmuch as they provide valuing agents with critical standpoints on their own emotions and inclinations, and even on the conceptions of good themselves, which make such agents potentially capable of self-revision and self-correction (in Susan Wolf's phrase).

But I am inclined to downplay the unrealistic suggestion of systematicity in Rawls' and Reid's discussions. A conception of "what is valuable in life" is an evaluative orientation, involving a concern to respond well to whatever practical circumstances throw our way. To lead a life is to be guided by such a conception. I shall remain agnostic here about how articulable the contents of evaluative orientations need to be.

(ii) the cultural presuppositions of Socrates' question…

(iii) where leading a life and having a life come apart. Culturally scripted lives….

(iv) *Leading a life and "narrativity"*

In a series of papers, John Fischer attempts to connect free action with narrativity. This is an important claim, if true. Although I am not here making any claims about the connection between acting freely, responsibility, and leading a life, plausibly there is some overlap. So if Fischer is right, then leading a life is, when it involves acting freely, a narrative activity.

---

[222] The other moral power definitive of persons is a sense of justice. This quoted passage continues: "Moreover, we must also include in such a conception a view of our relation to the world – religious, philosophical, or moral - by reference to which the value and significance of our ends and attachments are understood." John Rawls, "Justice as Fairness: Political not Metaphysical", as reprinted in *Collected Papers*, ed. Samuel Freeman, 398. (Harvard)

Fischer writes that "it is virtue of our exercising our capacity for …acting freely…that we are the kinds of creatures whose lives have an irreducible narrative dimension of value." "Introduction", *our Stories*, p. 14. In virtue of the exercise of this capacity, he adds, "we become authors of our life-stories".14

Fischer's starting point is the idea is that acting freely is a form of self-expression; indeed, that's on his account why so acting is pro tanto valuable. Self-expression is akin to a kind of artistic value, Fischer thinks. It gives narrative meaning to our lives. I want to assess this claim and say why I think it is indefensible.

To see how the gets to this claim, we have to begin with his intuitions about free actions as self-expressive.

Consider a remark from Sarah Broadie's book on Aristotle:

> "In voluntary action we pursue an objective which is before us and which figures as a good to us so far as we pursue it; but on another level we enact by our action, and thereby propound into public space, a conception of the kind of practical being that it is good (or at least all right) to be: a kind typified by pursuit of this kind of goal in this sort of way under such conditions." Ethics with Aristotle, 159

In "Responsibility and Self-expression", Fischer complains that the proposal in this passage runs afoul of weakness of will. We will not be able to find in such actions, even when one acts freely, an expression of what is good or worthy to pursue, or a "propounding" of a conception of practical wisdom. So we should look elsewhere for an account of what is expressed in free actions.

There are a number of things to be said on Broadie's behalf. But let's assume Fischer is right to think free actions are self-expressive, even akractic actions, and that Broadie's proposal fails to capture this. To this end, Fischer turns to the idea of narrative.

Fischer writes (Handout H): "I shall follow Velleman in contending that life has a narrative structure in the specific sense that the meanings and values of the parts of our lives are affected by their narrative relationships with other parts of our lives, and the welfare value of our lives as a whole are not simple additive functions of the values of the parts. In this sense, then, our lives are stories. And in performing an action at a given time, we can be understood as writing a sentence in the book of our lives." 290

So take Velleman's contrast between two scenarios: one in which one gives up on a difficult marriage and one in which one manages to work out the difficulties and achieve a deep and harmonious relationship. The latter scenario has a certain kind of value—a narrative value—inasmuch as the difficulties and struggles are redeemed by the successful efforts to stay together. So how does the possibility of values of this kind support Fischer's idea of free action as narrative authorship?

In contrast to Broadie: "what is expressed by an agent in acting [freely] is the meaning of the sentence of the book of his life. And this meaning is fixed in part by relationships to other sentences in this book, i.e., by the overall narrative structure of the life. In acting, an individual need not be 'propounding into public space' *any* sort of vision of the good or defensible life. Rather, his action writes part of the book of his life, and gets its meaning from its place in this story." 290-91

"Unlike …non-human animals , our lives are stories in a strict sense, and they can have a distinctive kind of meaning—narrative meaning." 167, "Stories and the Meaning of Life", in *Our Stories: Essays on Life, Death, and Free Will*. "When I act freely, I write a sentence in the story of my life; that is, the account of my life is strictly speaking a story (rather than a mere chronicle of events), and my life has a narrative dimension of value….[T]he value of our free action is a species of the value of artistic self-expression, *whatever that is*. I take it that artistic self-expression has value from the perspective of human flourishing or 'doing well'…." 167-8 (Fischer's emphasis)

Fischer emphasizes that the value in question here is not the value, aesthetic or otherwise, of the *product*. (Though he does say that "such free activity helps to endow our life-stories with the distinctive features of certain works of art."170.) The idea seems to be that acting freely has the value, "self-expressive", independent of its content, on which its moral and prudential value depends. Nor need it be, qua event-product, of any aesthetic value: it might be an ugly piece of vengeance. "When we act freely, we tell a story that is most naturally…evaluated in terms of moral and prudential considerations, even though the nature of the activity is artistic."169

I don't think it could possibly be true, at least not on the grounds given in these passages, that our lives are "strictly speaking" stories or that we are authors of these stories.

First, suppose that acting freely is in some sense writing a sentence in the story of one's life. It doesn't follow that "the account of my life is strictly speaking a story", for only a fraction of my life, and a fraction of my behavior, consists in free action, and thus the part "written by me" is only a fraction of my life, and the rest is written by no one. We are on this assumption at most entitled to speak of the part of my life that consists in my free actions as a story written by me.

But we are not entitled even to this much, it seems to me, for there is no reason to think that those "sentences" I "write" in acting freely compose in conjunction even part of a "narrative", or constitute (quasi)artistic self-expression. So even if I write some of the sentences that are part of my "life-story", it doesn't follow, and there is no reason to think, that in writing those sentences, what I am writing adds up to a story. (To see the point, suppose someone took three sentences from every one of my essays and wove them into a story. I wrote the sentences but was not a co-author of any part of the story.)

Furthermore, and most importantly, the metaphor of free actions writing sentences is surely extravagant. When we cash it in, all it comes to is the truth that in freely turning on the light, or getting a divorce, I freely make it true that I did these things. We are left with the humble if not trivial truth that our free actions make true certain sentences to which any account of our lives aspiring to truth must conform. **Now in the special case where I do something that redeems an earlier adversity, and do it for that reason, I freely affect the narrative value of my life. But this truth is miles away from the conclusion that free action as such write the books of our lives.**

(Note that there is nothing unique about free actions in this narrative respect. **One's emotions and beliefs can contribute complete a "narrative arc".)**

III Conclusion

Human beings are not just biological creatures; we are biographical creatures as well. We have lives not only in the sense that we are alive for a span of time, like plants and the other animals, but in the sense that we have ways of living. We have biographical powers in this sense: typically we not only think of ourselves as having a future and a past, but we have a guiding conception of how our future should go, and we care about how things have gone so far. Insofar as we are directed by a conception of how our lives should go, we can speak of *leading* a life. In this sense, the life one has can be damaged or destroyed without any distinct biological damage. Considerations about our biographical powers have led some writers to stress the importance of "narrative" and diachronic unity in human life. On this view, apart from the content of a person's life, there is a distinct value in having and fulfilling a life-plan, or carrying out something like a story. I have produced no arguments against narrative views, but I myself think that the importance of diachronicity is entirely a matter of the value of the substantive final ends that constitute the content of the lives people have. It is because having a worthwhile career, vocation, or discipline, or valuable attachments and devotions, inherently take time or belong to traditions and histories that diachronicity matters. It is because one has worthwhile ends that persisting in and realizing those ends is valuable, not because of any formal structural features of plan-fulfillment, or "narrative" coherence. The value of diachronicity is a side-effect of the value of certain activities and relationships, and value of unity of purpose, or of realizing one's life plans, for example, is not intrinsic but depends entirely on its enabling us to achieve independently valuable ends.