JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XII · NUMBER 1
September 2017

EDITORIAL

1 Ensuring a Future for Open-Access Publishing

ARTICLES

6 Ends to Means: A Probability-Raising Account of Means and the Weight of Reasons to Take Them
Matthew S. Bedke

29 Are All Normative Judgments Desire-Like?
Alex Gregory

56 Can There Be Government House Reasons For Action?
Hille Paakkunainen

DISCUSSIONS

94 Evolutionary Debunking Arguments and Our Shared Hatred of Pain
Ben Bramble

102 Interests, Wrongs, and the Injury Hypothesis
Richard Healey

110 Our Intuitions about the Experience Machine
Rach Cosker-Rowland

118 The Value of Caring: A Reply to Maguire
Jörg Löschke

127 Against Wolterstorff's Theistic Attempt to Ground Human Rights
David Redmond

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
http://www.jesp.org

The *Journal of Ethics and Social Philosophy* (ISSN 1559-3061) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge and is available to all readers under the CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 license. Funding for the journal has been made possible through the generous commitment of the Gould School of Law and the Dornsife College of Letters, Arts, and Sciences at the University of Southern California.

The *Journal of Ethics and Social Philosophy* aspires to be the leading venue for the best new work in the fields that it covers, and it is governed by a correspondingly high editorial standard. The journal welcomes submissions of articles in any of these and related fields of research. The journal is interested in work in the history of ethics that bears directly on topics of contemporary interest, but does not consider articles of purely historical interest. It is the view of the associate editors that the journal's high standard does not preclude publishing work that is critical in nature, provided that it is constructive, well-argued, current, and of sufficiently general interest.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.204

Editorial

ENSURING A FUTURE FOR OPEN-ACCESS PUBLISHING

THE *Journal of Ethics and Social Philosophy* (*JESP*) was founded in 2005 by Andrei Marmor, James Dreier, Julia Driver, and David Estlund, with the financial and technical support of the USC Gould School of Law and the USC Dana and David Dornsife College of Letters, Arts and Sciences, in which Marmor held joint appointments. The vision of *JESP*, then and now, is to provide a model for a directly university-funded, completely open-access publication at the very highest standard of peer review. *JESP*'s aspiration is to show how a journal can be at once one of the most well-run and most prestigious places to publish leading work in the fields of social, political, and legal philosophy and ethics, broadly construed, while doing so at zero cost to authors or readers, allowing for fast, easy publication with worldwide impact.

The rationale for the directly university-funded, open-access model is simple. The original rationale for the existence of journals was to disseminate research, and subscription costs were justified to support the expenses of dissemination, which for print journals were quite substantial on the margin. Over time, the peer-review structure of journals has come to play a second function—that of establishing a well-respected standard to help readers judge which articles are worth their attention and assist universities in evaluating the production of their scholars. Third, in an age of vast circulation of prepublished drafts, journal publication establishes a settled, final version of any piece of work for reference and discussion. University-funded open access is premised on the assumption that, given the advent of electronic publication and subsequent reduction in the marginal costs of dissemination, the first rationale for journals comes apart from the second and third. *JESP* is built on the understanding that any author can easily offer their own work for free access on their own website. Our role is not to provide access, but to set a standard of excellence that can call attention to work and establish a fixed record of that work.

Although the marginal costs of dissemination of any individual article is zero, *JESP* does bear the costs of evaluation of each submission and production of each accepted article. Every submitted paper is handled by a paid managing ed-

itor; every paper that goes out for review costs time and money to handle correspondence with possible referees; and every paper goes through copyediting and typesetting at the cost of both time and money. These costs don't go away in the era of electronic publishing, but the founders of *JESP* believed and still believe that bearing these costs falls under the central mission of modern research universities—the production and dissemination of knowledge. *JESP* has been blessed that USC administrators have shared this vision and are committed to permanent support of the journal.

The benefits of open-access publishing are I think made clear by *JESP*'s record. *JESP* published ten full articles in 2005–2006—the journal's first two years of publication—that have been downloaded more than 117,500 times—an average of more than one thousand per full year of publication, per article. Articles published so far in 2017 have been downloaded at an average rate of 278 times per month. Downloads are of course only one possible measure of reach, and most downloads vastly overrepresent actual readership. But the free availability of articles in *JESP* means that they may be read by scholars around the globe whose institutions cannot afford access to a subscription journal—and without imposing up-front costs on authors, many of whom are early-career scholars without access to research funds that could underwrite open-access fees at subscription journals. Even readers who have institutional access will in many cases find it easier to link directly to *JESP* from any device, without the need to log in through their institution or through a VPN client, which makes *JESP* articles more likely to be downloaded and read.

So it is no surprise that the ten full articles published by *JESP* in 2005–2006 have between them garnered more than 375 citations, according to Google Scholar—an average of 37.5 citations—and seven of them have been cited fifteen or more times. The five invited articles published as a symposium on Joseph Raz's "The Myth of Instrumental Rationality" have another sixty-seven citations; two of them have been cited at least thirty times. These are strong measures of impact for any philosophy journal over this period—compare that the first ten articles published in *The Journal of Philosophy* starting in April 2005 have totaled only 140 citations despite that journal's towering reputation established over more than a century. This suggests not only that *JESP*'s model provides an important service to the profession, but also that its model is in the best interests of authors who want their work to be read and engaged with.

I have been blessed to work with *JESP* in every possible role—as reader, author, referee, associate editor, and now editor-in-chief. In 2005, *JESP*'s founding year, I was invited to participate in *JESP*'s first symposium, on Joseph Raz's article, "Instrumental Rationality." I also submitted an original self-standing article to

JESP that year, "Cudworth on Normative Explanations," which I still believe to be the best paper that I have ever written. The following year I joined the USC faculty and became a regular referee for *JESP*, refereeing twelve papers over three to four years before being asked to serve as associate editor for discussion notes in 2009. In December 2014, I took over from Andrei Marmor as editor-in-chief in order to ensure continuity for *JESP*, as he planned for his departure from USC to a new position at Cornell University.

JESP's university-funded, open-access model comes with risks. So far as I know, *JESP* was the first leading, open-access philosophy journal to successfully go through a change in leadership, passing one of the first tests of a journal for institutional continuity that exceeds the personal investment of its founders. In addition, depending on direct university funding requires regularly seeking renewed commitments from university administrators, and the risks that this funding stream can depend too much on the involvement of particular individuals, with particular faculty appointments.

This year, *JESP* passed one more test to ensure continuity over time. Since its founding, it has been run on USC Gould School of Law servers using software originally developed by Gould IT staff. That software became obsolete over time, and with Marmor's departure and turnover in the Gould IT staff, it is no longer possible to maintain the journal using that system. So *JESP* has moved to a new online management system, the Open Journal System developed by the Public Knowledge Project, which facilitates effective management of open-access journals at a reasonable cost. Transferring *JESP* to the OJS has been labor-intensive but will ensure a stable future for the journal. It gives us greater functionality, and will be easier to maintain for the future. Along with this transition we have incorporated many other changes that will ensure the long-term accessibility of work published in *JESP*.

Among these changes, we are getting a new look—on the website and on the (electronically) printed page. When you visit *jesp.org*, you will now see a cleaner, easier-to-navigate site, with an updated *JESP* logo. You will find it easier to register as a user and volunteer to be a reviewer for *JESP*, find the content that you are looking for, and share *JESP* content on social media. *JESP* discussion notes, which were not an original feature of the journal and originally were not assigned to journal issues, have now been assigned to designated journal issues under a separate section for discussion notes, and past as well as future discussion notes can be found in this way, making them easier to cite following more standard conventions.

For *JESP* articles starting with volume XII, issue 1, Matthew Silverstein, Associate Professor of Philosophy at NYU Abu Dhabi, has volunteered his typesetting

skills and experience in order to produce a more polished look for the journal. Although *JESP* has always collected articles into volumes and issues, you will also find that we now take the division into volumes and issues more seriously; articles will carry continuous pagination across each volume as in traditional print journals; each journal issue is now designated with a unique jacket cover; and journal issues can be downloaded as unique PDF files, complete with a table of contents for that issue.

In summer 2016, *JESP* updated its editorial board, dropping a few inactive members and adding eight new members to increase energy and diversity of representation across geography, race, gender, field of expertise, career stage, and institutional affiliation. *JESP* is fortunate to have the continued support of its distinguished editorial board, and will continue to seek out new representatives over the coming years to more fully show distinction across the range of fields and institutions in which and from which we seek to publish the very best work. The full list of editorial board members can be found on the journal's website.

JESP continues to face challenges looking forward. Here I will single out just two. As I have already noted, in 2009 *JESP* added a discussion notes section in response to a perceived need for more places to publish short articles in philosophy. The need for this forum has been fully demonstrated by the response, as the number of discussion articles published in *JESP* now rivals the number of main articles. But we have found that the authors submitting discussion articles to *JESP* have been far less diverse in gender and areas of research than authors submitting main articles. One of *JESP*'s major objectives over the coming years is to appeal, particularly in the discussion notes section, to a much wider range of submitting authors.

A challenge we look forward to at *JESP* is securing more submissions in fields that fall under the label of "social philosophy," broadly construed, including philosophy of race, gender, and extra-political institutions like the family. *JESP* has published what I believe to be great work in these areas, including Anca Gheaus's article, "Gender Justice" (volume VI, issue 1); Jeremy Dunham and Holly Lawford-Smith's "Offsetting Race Privilege" (volume XI, issue 2); Luara Ferracioli's and Anca Gheaus's articles about children; and Joachum Wundisch and Vuko Andric's article on disability. But this falls far out of proportion with the role of "social philosophy" in the journal's title, and we hope to publish much more such work, and to find authors interested in publishing it in *JESP*.

Journal editorial work is a thankless task—virtually every day I put time into it and am reminded that it is by far my least favorite part of my job. It involves endless unpaid and unappreciated work, mostly having to judge the work of other philosophers. I have many constituents to please, and many ways of giving

them reasons not to be pleased. Authors expect prompt decisions and instructive justifications given for those decisions, referees expect their recommendations to be followed, and readers expect that work should be published in *JESP* only if it meets their own high standards. As editor-in-chief, I try to balance all of these demands and more, and depend at every stage on a hardworking and underappreciated team of associate editors, the enthusiasm of prospective authors for publishing their work in *JESP*, and the reliability and trustworthiness of hundreds of referees around the globe, many of whom I do not know, but who put substantial work into our unsolicited refereeing requests. It is not a perfect process, and it can go wrong in many ways, but we do our best, and each day, we try to do better—both at establishing a process that will produce the best results, and at respecting that process. We think we are doing well, but we would like to be the best, and we believe we can be. Thanks for your patience with us, and your continuing support for *JESP* and its mission.

Mark Schroeder
EDITOR-IN-CHIEF

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.207

ENDS TO MEANS

A PROBABILITY-RAISING ACCOUNT OF MEANS AND THE WEIGHT OF REASONS TO TAKE THEM

Matthew S. Bedke

LET US SAY THAT YOUR ENDS are *whatever* you have ultimate (underived) reason to do or to bring about. This leaves open whether your ends are stance dependent and so "given" to you by your contingent desires, your nature as a rational being, your self-identity, etc., or whether at least some of your ends are not stance dependent and so not "given" to you in any of these ways.

So understood, it is uncontroversial that you have reason to take the means to your ends. More specifically, some reason to do or bring about an end is going to transmit to reason to take the means. Using this as a point of departure, this paper considers *what it is* to be a means to an end and *how much* reason transmits from an end to its means. The theory on offer is a probability-raising theory that says roughly this: an action is a means to an end just in case it raises the probability of the end relative to *the worst one could do*—i.e., relative to that action that would make the end least probable. And the amount of reason transmitted from an end to a given means is a function of the degree to which it raises the probability of the end relative to the worst one could do.

Some recent literature addresses these issues, though given the importance of means-ends relations and the ascendance of reason-based analyses of normativity the issues remain relatively underexplored.[1] My own exploration will proceed

1 The account here is meant to improve upon and replace the Transmission Principle in Bedke, "The Iffiest Oughts." For similar probability-raising accounts, see Kolodny, "Instrumental Reasons"; and Stegenga, "Probabilizing the End." For a very different approach, see Horty, "Reasons as Defaults." For the related issue of what it is to promote the satisfaction of one's desires, see Lin, "Simple Probabilistic Promotion"; Sharadin, "Checking the Neighborhood" and "Problems for Pure Probabilism about Promotion (and a Disjunctive Alternative)"; Coates, "An Actual-Sequence Theory of Promotion"; Schroeder, *Slaves of the Passions*; and Finlay, *Confusion of Tongues*. For a skeptical take on the probability-raising approach to promotion, see Behrends and DiPaolo, "Finlay and Schroeder on Promoting a Desire" and "Probabilistic Promotion Revisited." The present project is related to those discussions of means-ends reasons and reasoning that focus on necessary means (see, e.g., Bratman, "Intention, Practical Rationality, and Self-Governance," 424; Darwall, *Impartial*

by generating plausible proposals within a certain framework and "Chisholming" to a correct analysis. Though I do not offer a comprehensive compare and contrast with other approaches in the literature, along the way I indicate some of my reservations about some of those approaches.

1. WHY PURE PROBABILITY RAISING?

Before I begin, let me briefly address some objections to a pure probability-raising approach, which aspires to offer necessary and sufficient conditions for being a means and to provide resources for modeling reason transmission. I hope I can do this in a helpful way before nailing down the best version of a probability-raising theory and answering some intramural questions within that approach.

One big-picture issue is whether probability raising of any sort is necessary for being a means. The crucial question here is whether there are means to an end that do not raise the probability of the end. In that vein, consider the following. Ann is golfing and one of her ends is to get a hole in one. If she slices the ball, intuitively that reduces the probability she will get a hole in one relative to some relevant baseline (e.g., hitting a clean shot).[2] But on this occasion, Ann gets lucky. Though her tee shot slices the ball, she hits it *just so* and the result is a hole in one. Arguably, this is a case in which an action (slicing the ball) is a means to an end (getting a hole in one) by virtue of causing it (/bringing it about) and despite the fact that it lowered the probability of that outcome relative to a relevant baseline.[3] If so, probability raising is not necessary, and being an action that causes, brings about or constitutes an end is sufficient for being a means.

Cases like this raise tricky issues. One question is how to demarcate an action so that we might ask of *it* whether it is a cause or a probability raiser or a means to an end. For instance, are we to consider whether *slicing the ball* is a means to the hole in one, or *slicing the ball just so* is a means to the hole in one? It looks like the coarse-grained action of slicing (any which way, as it were) has less claim to being a cause of the end than the more fine-grained action of slicing it just so, for there are lots of ways in which the coarse-grained action can play out and very

Reason, 16; Schroeder, "Means-End Coherence, Stringency, and Subjective Reasons," 245) or sufficient means (see, e.g., Raz, "Simple Probabilistic Promotion," 9, and "The Myth of Instrumental Rationality," 5–6), but the scope here is intended to be broader.

2 Articulating the relevant baseline is one of the intramural issues that needs to be sorted. But I think the case raises a nice structural question that we can address without first settling on a baseline.

3 The example is adapted from discussions of causation as probability raising in Hitchcock, "Do All and Only Causes Raise the Probabilities of Effects?" and Suppes, *A Probabilistic Theory of Causality*, 41.

few of them result in a hole in one, whereas the fine-grained action leaves little room for ways of doing it that do not result in a hole in one. So let us first consider a specification of the case that focuses on the fine-grained action.

Is the fine-grained action a probability raiser? Arguably, yes. Indeed, there are plausible, objective theories of probability that will assign conditional probability one to an effect given its cause.[4] More generally, there are ways of theorizing about probability raising that can be applied to the golfer's case, which presents an interesting theoretical choice: either plug in a theory of probability that makes slicing the ball *just so* (and other acts that cause/bring about/constitute some end) a probability raiser and thereby defuse the counterexample, or choose a theory of probability under which the fine-grained action fails to be a probability raiser and thereby give the counterexample some teeth. Faced with this choice, the probability-raising approach looks like an attractive, simple, theoretical option flexible enough to handle the case, and not necessarily an approach that succumbs to a clear, devastating counterexample.

On the other hand, it is not clear to me that we should be focusing on the fine-grained action to begin with. For our purposes, a means needs to be an action over which we exercise some *agency* and over which we have some *control*. If Ann is like most of us, she does not have the requisite control over how she slices a golf ball to be able to slice it *just so*. The fine-grained specification of the case does not feature an item eligible for being a means to an end and once again we defuse the counterexample. Note that the problem here is not one of outcome luck. The problem is whether Ann has the agentive ability to produce the requisite intentions and motor functions—the ones that guide slicing the ball *just so*, resulting in a hole in one—rather than nearby intentions and motor functions.

So let us consider a specification of the case that focuses on the coarse-grained action. Though slicing the ball (any which way, as it were) plausibly lowers the probability of the end relative to some relevant baseline, we have already noted that it has a much weaker claim to causing the end. Insofar as we undermine the case for causation, we also undermine the intuitive sense that this action is really a means to the end. Once again, we do not have a clear counterexample—we do not have a clear case of a means that is not a probability raiser. At the very least, these considerations make it difficult to craft a case of an action i) over which

4 I will not have much more to say about which theory of probability we should use, for it could be that different theories of probability serve different contexts. It could be, for example, that some theory of subjective probability is best for a first-person deliberative context, or for assigning blame or praise, or grading character, while some objective theory is best for giving advice or evaluative assessment without blame. Having said that, I have written this paper with some notion of objective probability in mind.

one has agentic control (ii) that causes, brings about or constitutes an end while (iii) resisting a probability-raising analysis.

Suppose we can craft the relevant counterexample that calls into question the necessity of probability raising. Where would that leave us? I think it would leave us with a disjunctive theory of means and their weights. We would say that one set of conditions is sufficient to be a means—causing, bringing about or constituting an end—but we would need a different set of conditions for those means that do *not* cause or bring about or constitute an end. We would also need some account of how reasons transmit in the two cases, and it too might be disjunctive. By contrast, a probability-raising approach promises a unified account of what it is to be a means (probability raising) that dovetails with a unified account of weight transmission (probability raising). So even if we could generate a nice counterexample, we would need to decide between the strength of the counterexample and the relative simplicity and elegance of the probability-raising approach.[5] While adjudication of that issue must await the convincing counterexamples to probability raising and development of the best probability-raising theory (something I try to do below), I hope I have said enough to justify an approach that takes probability raising to be a necessary condition on being a means.

The other big picture issue is whether probability raising is sufficient for being a means to an end. Kolodny offers a case in which a boxer's end is to land a punch and the boxer has a "tell" each and every time he is about to throw a punch.[6] Intuitively, a tell like this seems to be more of an *indicator* or a *pre-signal* that a punch is coming and less like a means to landing a punch *even if* the prob-

5 Lin ("Simple Probabilistic Promotion") has a different argument against causation as a sufficient condition. His argument focuses on cases in which one can cause an end by doing an action, like pressing a button, but the end will occur even if one does nothing (e.g., someone else ensures that the end is brought about even if one does nothing). It seems that *any* amount of reason to do nothing (e.g., pressing the button will produce a mild electric shock) will outweigh the reason to press the button, for the end will occur regardless. This, Lin argues, shows that one does not have *any* reason to press the button, otherwise it would have some weight capable of outweighing some small reason to do nothing. I am not convinced by the argument. One who thinks causing an end (/bringing it about/constituting it) is sufficient to be a means could say that, in these cases, the end transmits *just as much reason to do nothing as it transmits reason to press the button*. After all, the end occurs regardless of what one does. If so, we predict that any *other* reason to do nothing (e.g., the electric shock) will *combine* with the end-given reason to do nothing, and these together will outweigh the end-given reason to press the button. That said, like Lin I reject causing an end as sufficient for being a means, but based on the parsimony considerations in the main text.

6 Kolodny, "Instrumental Reasons."

ability of landing a punch given the tell is greater when compared to a relevant baseline.

Once more, cases like this raise tricky issues. First, we should reiterate a point made above: items eligible to be a means are actions that involve agency and control. So we should rule out interpretations of the boxer's case in which the tell lacks agentive control. Having said that, there could be interpretations of this case or cases like it in which the tell is subject to agentive control, but where it still seems to fall short as a means. So we should further clarify, second, that an action is not a probability *raiser* for an end simply because the conditional probability of the end given the action is higher than some relevant baseline. To be a probability raiser the action must *make* the probability of the end higher than some relevant baseline—the probability of the end has to be higher at least partly *in virtue of* the action and its relations to the relevant baseline and the end.[7] At least, this is a feature of the present approach. Even if the boxer's tell is subject to agentive control, intuitively it is not helping to *make* the probability of landing a punch higher (the throwing of the punch does that) even if the probability of landing the punch is in fact higher given the tell compared to some relevant baseline. Since it is not a probability raiser, *a fortiori* this is not a case of a probability raiser that fails to be a means.

We now have two responses to cases that appear to be counterexamples to probability raising as a sufficient condition on being a means: make sure the case concerns actions in the relevant sense and make sure the action helps to make the probability of some end higher. Given these constraints, it is more difficult to generate a counterexample to sufficiency.

Kolodny has a different solution to the boxer's case.[8] He says that a means must be an action such that the probability that the action *helps to bring about the end*, conditional on the action, is positive. And he thinks the boxer's tell does not help to bring about the end of landing a punch. This solution differs from mine at least nominally, for Kolodny has a bringing-about relation between an action

7 Cf. Lin, "Simple Probabilistic Promotion." Lin goes on to characterize this in-virtue-of relation as causal, but I think that is unduly restrictive. Side note: this clarification about probability raisers has interesting applications in decision theory. In Newcomb's paradox, for example, choosing one box arguably does not make the probability that there is $1 million in the opaque box higher even if the Pr ($1 million in opaque box | choose one box) is higher than the relevant baseline. Arguably, it is the categorical basis for the tendency to choose one box (detectable by the predictor some time before the present choice) that helps to raise the probability that there is $1 million in the opaque box. So choosing one box might not be the best means to getting the most money under a probability-raising theory of means.

8 See Kolodny, "Instrumental Reasons."

and an end, whereas I have a making-it-the-case relation between an action and a conditional probability. I think the following case might help bring Kolodny's constraint into focus and show why I think it is mistaken.

Suppose a rogue state has launched a nuclear warhead and the end is to neutralize the threat. A missile defense system is in place, and one intercept has already been fired with .5 probability of neutralizing the threat. Let us assume this .5 probability is the relevant baseline. You have three available actions and you can only do one. You can do nothing, leaving the probability of neutralizing the threat at .5. You can fire another intercept that raises the probability of neutralizing the threat to .75, for if the first misses, there is .5 probability that the second hits. Finally, you can press a button that sends an electromagnetic pulse through the atmosphere that raises the probability of neutralizing the threat to .7, for the signal has .4 probability of deactivating the nuclear warhead before the first intercept comes close, combined with even odds that the first intercept works if the pulse does not.

It seems that you have most derived, means-based reason to fire the second intercept rather than push the button that sends a pulse or do nothing, for it raises the probability of the end from .5 to .75, whereas pressing the button only raises it to .7 and doing nothing leaves it at .5. But notice that, given that you fire the second intercept, it only has a probability of .25 of *helping to bring about* the end, for it can only do that if the first intercept misses, in which case it has even odds of intercepting. By contrast, the probability that pressing the button helps to bring about the end is .4, for it will work (or not) before the first intercept arrives. In light of this, it looks like Kolodny's condition will deliver greater derived reason to press the button than to fire the second intercept even though firing the second intercept makes the end more likely than pressing the button. I think that is the wrong result.

To be fair, Kolodny wants the "helping to bring about" clause to be read broadly, and maybe it can be read broadly enough so that his condition of helping to bring about an end is extensionally equivalent to my condition of *making* the end more probable. If so, I think the probability-raising condition I suggest is a clearer way to identify that extension, and the better candidate for what it is to be a means.

Let me give one more schematic case to support probability raising as sufficient for being a means. Suppose for the sake of argument that we add some extra necessary condition such that to be a means is to be a probability raiser *plus some x-factor*. This would rule out those actions that lack the x-factor even if they are really excellent probability raisers. Now consider a case in which Alice has

two ends, e_x and e_y, that are probabilistically independent of one another, and two available actions, a_x and a_y. Further:

a_x raises the probability of e_x by .5, and the x-factor is present.
a_y raises the probability of e_x by .5, and the x-factor is present.
a_x raises the probability of e_y by .99, and the x-factor is absent.
a_y raises the probability of e_y by .5, and the x-factor is present.

If the x-factor is a necessary condition, overall a_y is a means to *both* ends, while a_x is a means to only *one* end. If we use probability raising as a guide to the weight of transmitted reasons, Alice would have most reason to do a_y, for it raises the probability of each end by .5. Meanwhile, a_x lacks the x-factor vis-à-vis e_y, so it is not a means to that end and inherits no transmitted reason weight from that end even though it makes that end almost certain to obtain, and far more likely to obtain than does a_y.

That is the wrong result. If ends are what ultimately matter, it is hard to see why a_y would derivatively matter more than a_x. After all, a_x *makes* the things that ultimately matter more likely to occur compared to a_y. This is not a conclusive criticism of all x-factor conditions, of course, and there is the possibility of other objections against probability raising as sufficient for being a means. But I hope I have said enough to justify my focus on pure probability-raising accounts. That is the task for the rest of the paper.

2. FRAMEWORK

In what follows, I will work within a certain theoretical apparatus. Let us suppose that, for a given agent, S, and context of choice, C, there is a finite set of actions, $A = \{a_0, a_1, \ldots, a_n\}$, where these actions are (i) available to S in C, (ii) jointly exhaustive, and (iii) mutually incompatible with one another.[9] I will not spell out what it is to be an available action (though I have said the actions need to involve agential activity over which one exercises some control), or whether there is a privileged set of exhaustive and incompatible actions to consider for a given agent in a given context. Those are debatable issues and plausible theories on those fronts can be plugged into the general framework here.

Suppose also that for an agent and a context there is a unique, finite set of ends, $E = \{e_0, e_1, \ldots, e_n\}$. And suppose that for any end, $e_i \in E$, its ultimate weight can be modeled by assigning it a positive number, g_{ei}, on some cardinal scale of weights (possibly different scales for different ends).

9 The approach here is consistent with contrastivism about reasons (see Snedegar, "Contrastive Reasons and Promotion" and "Reason Claims and Contrastivism about Reasons").

Relative to an agent (S), context of choice (C), set of available actions (A) and set of ends (E), we want to analyze what it would be for some $a_i \in A$ to be a means to some $e_i \in E$ and how much weight transmits from a given end to its means. The point of specifying the analysis relative to some S, C, A, and E is to model as best we can what agents are faced with when they act, and what considerations other parties can draw on for evaluating the actions of others. We want to do so in a way that allows sensible comparisons, trading off and aggregation of what one has reason to do without shifting the relevant context of choice and options to choose from in the process.

I will proceed by considering various *prima facie* plausible baselines for probability raising. An appendix will discuss some related matters, including what it is to transmit a reason *against* an action, and how to approach the nettlesome problems of aggregation to oughts, incommensurability, and defeat.

3. Refraining as an Independent Action

For an action $a_i \in A$ at a context of choice C the obvious baseline to consider is the case of not-a_i. Does a_i raise the probability of some end relative to not-a_i? There are a number of ways to regiment the idea. Let us consider them one by one.

First, one might appeal to *refraining from a_i-ing* as its own independent action in the set of available actions. The view I have in mind would deviate from some standard ways of characterizing what it is to refrain. On one standard view, one can refrain from doing an action *while* doing some other action. If so, I can refrain from staying home at the same time I do a distinct action, like going to the park. Of course, such a view would not allow us to construct a set of actions A with both *go to the park* and *refrain from staying home* in the set, for even if these are two available actions they are not incompatible, and incompatibility is a constraint on actions in A. On another standard view, *what it is* to refrain from one action just is to do some incompatible action. So by going to the park I refrain from staying home. Again, this view does not allow us to construct a set of actions with both *go to the park* and *refrain from staying home* in the set, for they are not incompatible.

Set those characterizations of refraining to one side for a moment. I want to consider a nonstandard view in which refraining from *a*-ing is an available action that is not only incompatible with *a*-ing, but is also incompatible with other actions incompatible with *a*-ing. You might doubt that this is the best view of refraining, but let us grant that it is intelligible and let us consider the view that an action is a means to an end just in case it raises the probability of the end relative to refraining so understood. More generally, for a given a_i the relevant baseline

would be (i) refraining from a_i, if a_i is *not* a refraining, or (ii) the non-refraining counterpart to a_i if it *is* a refraining. If we let "$r(a_i)$" be the symmetric refraining function on a_i, and use this for the relevant baseline, the proposal is this: for some $a_i \in A$ to be a means to some $e_i \in E$ is for:

1. $a_i \neq e_i$,[10] and
2. Partly in virtue of a_i, $\Pr(e_i \mid a_i) > \Pr(e_i \mid r(a_i))$

The means-based weight for a_i, w_{ai}, which is transmitted from the end-based weight for e_i, g_{ei}, would be proportional to the degree to which a_i raises the probability of e_i from the baseline as follows:

3. $w_{ai} = g_{ei} \times (\Pr(e_i \mid a_i) - \Pr(e_i \mid r(a_i)))$

On this analysis, if finishing the paper is among my ends, and the available actions are {stay home, refrain from staying home, go to the park, refrain from going to the park, go to the movies, refrain from going to the movies}, then staying home is a means to finishing this paper just in case that action makes greater the Pr(I finish this paper | I stay home) compared to the Pr(I finish this paper | I refrain from staying home). The big problem with this view is that each of staying home and refraining from staying home might make it highly likely that I finish the paper. Suppose, for instance, the following probabilities:

Pr (I finish the paper | I stay home) = .75
Pr (I finish the paper | I refrain from staying home) = .75

Either way, I am very likely to finish the paper. And because staying home does not raise the probability of finishing the paper relative to refraining from doing so I would have no means-based reason to stay home if we take refraining to be the relevant baseline. This is particularly odd when we compare some other available actions that make finishing the paper much less likely. For the other actions available to me, suppose the following conditional probabilities:[11]

Pr (I finish the paper | I stay home) = .75
Pr (I finish the paper | I refrain from staying home) = .75
Pr (I finish the paper | I go to the park) = .5
Pr (I finish the paper | I refrain from going to the park) = .4

10 An end cannot be a means to itself on pain of double counting the weight of reasons that support it, once as an ultimate reason and once as a transmitted reason.

11 Henceforth, let it be understood that the examples involving higher conditional probabilities compared to a baseline include the requisite relation of *making* the end more probable unless otherwise specified.

If we take refraining as the baseline, my end of finishing the paper gives me no reason to stay home but some reason to go to the park. Staying home would not be a means at all whereas going to the park would. These are the wrong results. When defining means and the weights of reasons supporting them, we do not want the two most successful actions (probability-wise) to be baselines for each other, nor do we want any two poorer actions to be baselines for each other.

4. DOING SOME ALTERNATIVE

To get away from confounding pair-wise comparisons, suppose we take the baseline for any given $a_i \in A$ to be *doing some alternative* to a_i, in which an alternative to a_i is an action in the set A incompatible with a_i (we saw that this is one view of what it is to refrain from a_i, and it is also one interpretation of the referent of "not- a_i"). On this view, for an action to be a means is for that action to raise the probability of an end relative to *doing something else*, and the amount of reason transmitted is proportional to how much that probability is raised. More formally, for some $a_i \in A$ to be a means to some $e_i \in E$ is for:

1. $a_i \neq e_i$, and
2. Partly in virtue of a_i, $\Pr(e_i \mid a_i) > (\Pr(e_i \mid a_0) + \Pr(e_i \mid a_1) + \ldots + \Pr(e_i \mid a_n) - \Pr(e_i \mid a_i)) \div n$

The means-based weight, w_{ai}, relative to ultimate reason weight, g_{ei}, is this:

3. $w_{ai} = g_{ei} \times (\Pr(e_i \mid a_i) - ((\Pr(e_i \mid a_0) + \Pr(e_i \mid a_1) + \ldots + \Pr(e_i \mid a_n) - \Pr(e_i \mid a_i)) \div \mathrm{n})$

This gets away from confounding pair-wise comparisons, so that is good. The problem is that this baseline makes a given *a*'s status as a means, and the weight of reason transmitted to it too sensitive to the presence of *better* means. To see why, consider a set of available actions, {stay home, go to the park, go to the movies}, and suppose the following probabilities:

Pr (I finish the paper | I go to the park) = .5
Pr (I finish the paper | I go to the movies) = .25
Pr (I finish the paper | I stay home) = .75
Pr (I finish the paper | I go to the movies or stay home) = .5

Under the analysis, going to the park would not be a means to finishing the paper, for it does not raise the probability of finishing the paper relative to doing something else (going to the movies or staying home). That seems strange given that one of the actions available to me makes finishing the paper much less probable.

In addition, when we use doing something else as a baseline, small changes in the best available action *could* turn something that is not a means into a means. To see this, contrast the above example with the following:

Pr (I finish the paper | I go to the park) = .5
Pr (I finish the paper | I go to the movies) = .25
Pr (I finish the paper | I stay home) = .74
Pr (I finish the paper | I go to the movies or stay home) = .495

In this scenario, going to the park *is* a means with some weight. But if the decrease in Pr (I finish the paper | I go to the movies or stay home) is due to the decrease in Pr (I finish the paper | I stay home), we have a pair of cases in which the status of going to the park as a means depends on how effective my *best option* is at probabilizing the end.[12] Again, that is the wrong result. When we use *doing something else* as the relevant baseline, it accords too much influence to the best available actions.

5. BEFORE VS. AFTER

A similar problem afflicts probability-raising accounts in which the relevant baseline is the probability of the end *prior to acting*.[13] Consider a case in which the probability that I finish the paper is very high simply because it is very likely that I will stay home and very likely that I will finish the paper given that I stay home. Indeed, suppose that this makes the probability of finishing the paper much higher than the probability of finishing the paper given that I go to the park. In that case, going to the park is not a means at all, even if it makes the probability of finishing the paper much higher than something else I could do, like going to the movies. That is the wrong result. Using the probability of the end prior to acting as a baseline gives too much influence to my best option and how likely I am to perform it.

6. WHAT WOULD OTHERWISE HAPPEN

One might seek a baseline in whatever happens in the nearest possible world where a_i fails to occur (yet another possible referent of "not-a_i"). Let us then condition on $\sim a_i$, where this operation is understood to take us to the nearest

12 It is not generally the case that Pr $(e \mid (a_1$ or $a_2)) = [\Pr(e \mid a_1) + \Pr(e \mid a_2)] \div 2$. But there are cases where this occurs.

13 Lin ("Simple Probabilistic Promotion") has this probability-raising theory of promotion. He acknowledges the problem I point out in the text.

possible world(s) where a_i does not occur, and ask after the probability of e_i in that/those world(s).[14] On this view, for some $a_i \in A$ to be a means to some $e_i \in E$ is for:

1. $a_i \neq e_i$, and
2. Partly in virtue of a_i, $\Pr(e_i \mid a_i) > \Pr(e_i \mid {\sim}a_i)$

The means-based weight, w_{ai}, relative to ultimate reason weight, g_{ei}, is:

3. $w_{ai} = g_{ei} \times (\Pr(e_i \mid a_i) - \Pr(e_i \mid {\sim}a_i))$

A special case here deserves special attention—the case in which ${\sim}a_i$ takes us to a world where one *fails to act at all*; i.e., *C* and *A* are held fixed and *S* fails to do any action in *A*. I think it would be a mistake to appeal to a nonaction for a baseline. Suppose the worst thing I could *do* vis-à-vis finishing the paper is to go to the movies, where Pr (I finish this paper | I go to the movies) = .25, as before, and *no other available action makes finishing the paper less likely*. It just so happens that I will go to the movies, but if I were not to go to the movies, I would pass out, where Pr (I finish this paper | I pass out) = .15 (I might be able to finish it after I come to, but I will be distracted by medical concerns). It follows on the present analysis that going to the movies is a means to finishing the paper. For it probabilizes the end relative to what happens in the nearest world in which I do not go to the movies.

That might seem counterintuitive by itself. How could the worst thing I could do be a means to my end? A deeper problem is that allowing for nonaction baselines frustrates the first-personal goal of helping to decide what to do and the second- and third-personal goals of evaluating what *is done* against what *could have been done*. I cannot choose an action over a nonaction—I cannot decide to go to the movies rather than pass out. And it seems illegitimate to take into account nonactions when evaluating what is done. It is illegitimate, for instance, to positively evaluate going to the movies to some degree because of the possibility that I pass out. So I think it is inappropriate to go outside the set of available actions to set baselines.

The point might be easier to see with a case in which nonaction makes achieving an end most likely. So consider a context in which my end is that my mother is taken care of. I could do a lot in this regard. I could try to take care of her myself, I could make various arrangements, etc. Other things would frustrate this end, like going to the casino. But suppose that, for each available action, it turns

14 Finlay ("The Reasons that Matter," 8n19) arguably has this view. In *Confusion of Tongues*, 38–46, he seems to have a different view.

out that in the nearest possible world in which it does not occur I pass out.[15] It also turns out that in those nearest possible worlds others will certainly ensure my mother is taken care of. In this case, every action I could do actually reduces the probability of the end relative to the $\sim a_i$ baseline, so none of them are means. That is, for all $a_i \in A$, $\Pr(e_i \mid a_i) < \Pr(e_i \mid \sim a_i)$. On such a view, there is no reason to choose an action in virtue of being a means to my end. But surely some of the things I can do better achieve my end than others—arranging for my mother's care is better than going to the casino. To extend the point made above, I cannot use the results of the analysis to make any decisions, and those wishing to evaluate my conduct cannot appeal to the fact that things go best when I pass out.

It is tempting to say that the best thing to do is that which will bring about my passing out. But that is a very particular kind of case. The hard case to focus on is one where nothing I can do probabilizes passing out, and each thing I can do fails to probabilize the end relative to what happens if I do not do that thing (where we stipulate that, for each action, if I do not do it I pass out). This kind of case helps show why *not acting* cannot be a relevant baseline.

All this can be avoided by allowing "$\sim a_i$" to range only over the nearest possible world(s) in which one *performs an action* other than a_i.[16] This is only a partial fix. It might help meet the deliberative and evaluative concerns, but it still misidentifies means. If in the nearest possible world in which I perform an action other than a_i I perform the action that makes the end most probable, this should not thereby make a_i a non-means. It should be highly relevant whether I can do things other than a_i that make my end even less likely than does a_i. As with using the performance of some alternative action as a baseline, this one is too sensitive to better means and not sensitive enough to worse ones.

Note also that, like the refraining function above, subjunctives do not pick out a unique baseline for all actions in A. There is therefore potential to assign means-based weights in a way that skews the comparison of reasons for the actions in A. In the paper-writing case, suppose a set of available actions A = {stay home, go to the coffee shop, go to the park, go to the movies}, in which going to the coffee shop and staying home make finishing the paper equally likely and more likely than anything else I could do. More specifically, suppose:

15 This seems to entail that I will actually pass out, in which case you might think that the actions in A are not available after all. But we should not take what actually happens to limit what can happen.

16 This is similar to the baseline Jackson and Pargetter fix when the question is whether to do a_i ("Oughts, Options, and Actualism," 246), although, in effect, they think this context of choice identifies the set of available actions with the set $\{a_i, \sim a_i\}$. I will revisit this issue below.

Pr (I finish the paper | I stay home) = .75
Pr (I finish the paper | I go to the coffee shop) = .75
Pr (I finish the paper | I go to the park) = .5
Pr (I finish the paper | I go to the movies) = .25

Further, I *will* go to the coffee shop. If I do not go to the coffee shop, I go to the movies. If I do not stay home, I go to the coffee shop. On this fact pattern, staying home is not even a means, and no reason transmits to it, while going to the coffee shop is a very good means, supported by weighty means-based reason. That is the wrong result.[17] The problem is a result of shifting baselines.

7. THE STATUS QUO

Could the baseline be the *status quo*, such that an action is a means to an end if it raises the probability of that end relative to the status quo?[18] Though it is not entirely clear what this baseline amounts to, it sounds like the analysis would make it impossible to have among *A* the action of *carrying on with the status quo*, and for that action to be a good means to one's end. For carrying on with the status quo would not raise the probability of an end relative to the status quo. That is the wrong result. Sometimes we are already in the process of taking a means to one of our ends. Indeed, sometimes we are already taking the best means and we should carry on.[19]

8. ANY POSITIVE CONDITIONAL PROBABILITY

Now consider the zero baseline: to be a means to an end just is to make that end probable *to any positive degree*. More formally, for some $a_i \in A$ to be a means to some $e_i \in E$ is for:

17 Cf. Behrends and DiPaolo, "Finlay and Schroeder on Promoting a Desire," 2–3

18 Mark Schroeder defends the view that one has reason to do that which *promotes* the object of any of one's desires, where promotion is understood probabilistically but where the weight of reason is not proportional to the degree of promotion (*Slaves of the Passions*, 113). He characterizes the baseline for promotion both in terms of *doing nothing* and the *status quo*. To be fair, his primary aim in that work is not to settle the details of the promotion relation, but to press a defensive strategy for Humean theories of reasons.

19 Cf. Evers, "Humean Agent-Neutral Reasons?" 60; and Behrends and DiPaolo, "Finlay and Schroeder on Promoting a Desire," 4–5. In their criticism, Evers and Behrends and DiPaolo focus on the "do nothing" baseline, while in this section I focus on the "status quo" baseline. It is not at all clear that the two baselines are equivalent, for the status quo might not involve doing nothing.

1. $a_i \neq e_i$, and
2. Partly in virtue of a_i, $\Pr(e_i \mid a_i) > 0$

The means-based weight, w_{ai}, relative to ultimate reason weight, g_{ei}, is:

3. $w_{ai} = g_{ei} \times \Pr(e_i \mid a_i)$

This analysis is nice and simple.[20] By including the "in virtue of" clause in 2, we avoid some very natural concerns. It is natural to wonder, for example, about the special case in which $\Pr(e_i) = 1$. Does this analysis make all actions ($\neq e_i$) in any set A means, supported by as much reason as the end itself? No, thanks to the "in virtue of" clause. If the $\Pr(e_i) = 1$ *because* of some action or actions in A, then those actions that are making the probability go to 1 will be probability raisers and they will get some weight transmitted to them. But if no action of yours is helping to make it the case that $\Pr(e_i) = 1$, then they will not count as means and will receive no weight from the end. Whatever the probability of e_i is, the question for this analysis is whether some action of yours would make its probability some amount greater than zero. Then and only then is it a probability raiser.

The troublesome case for this proposal is a case in which some action of yours makes the probability of some end greater than zero, but the action makes the end *as unlikely as possible*. If a_i is the absolute worst thing I could do vis-à-vis e_i (probabilistically speaking), but doing so still yields a positive probability for e_i, the analysis counts a_i as a means to e_i, supported by some weight. Most pointedly, a_i would be a means supported by some weight even if every other action available to you makes the end much more likely. Suppose, for example, that the prior probability of finishing the paper is .5, and of my three available actions (stay home, go to the park, go to the movies), going to the movies drops the probability to .1, while going to the park keeps it at .5 and staying home raises it to .9. This is a case in which going to the movies is a difference maker, probability wise, but in a bad way. Calling it a means seems inapt—if I make finishing the paper as unlikely as I can, I am not taking a means to finishing it. To solve this concern, some relevant baseline other than zero is called for.[21]

20 And I think it is the one used by Kolodny ("Instrumental Reasons"). This is as good a place as any to point out that Kolodny needs a clause in his transmission principle to block iterations of the principle that would generate reasons that do not flow from ultimate ends. We handle the problem by defining means and analyzing weights directly in terms of those things we have *ultimate* reason to do.

21 As one reviewer pointed out, one advantage of the zero baseline is that it makes the status as a means independent of the set of available actions, A, that we construct. By contrast, my favored view below is highly "menu dependent": whether an action is a means and how much reason gets transmitted to it depends on the set of actions, A, in which it is found, though I

9. THE WORST ONE COULD DO

To sum up the lessons from above, we are in need of a unique baseline for all actions in A, one that is itself an action and that takes the availability of worse actions vis-à-vis an end to be more probative than the availability of better actions, but which does not count any action with positive conditional probability for an end as a means. So let us consider the view that an action is a means to your end insofar as it raises the probability of the end relative to the worst you could do. More formally, let us characterize the worst you can do like this: for a given $e_i \in E$, the worst you can do is that $a_u \in A$ such that $\Pr(e_i \mid a_u) \leq \Pr(e_i \mid a_j)$, for all $a_j \in A$. Then for some $a_i \in A$ to be a means to some $e_i \in E$ is for:

1. $a_i \neq e_i$, and
2. Partly in virtue of a_i, $\Pr(e_i \mid a_i) > \Pr(e_i \mid a_u)$

The means-based weight, w_{ai}, relative to ultimate reason weight, g_{ei}, is:

3. $w_{ai} = g_{ei} \times (\Pr(e_i \mid a_i) - \Pr(e_i \mid a_u))$

I think this is the best probability-raising account of means and the weight of reasons supporting them. It meets the desiderata spelled out above. It is an action, it makes the weight of reasons to take means insensitive to better means and it does not necessarily count actions with any positive conditional probability as means.

The present proposal has some similarities with actualism, defended by Jackson and Pargetter ("Oughts, Options, and Actualism"). To compare the views, let us consider the paper-writing case again, where A is {stay home, go to the park, go to the movies}, and suppose that if I go to the park I will get injured, spend some time in the hospital and likely finish my paper there, whereas if I stay home I stand a good chance of playing video games instead of finishing the paper. The probabilities are:

Pr (I finish the paper | I stay home) = .5
Pr (I finish the paper | I go to the park) = .75
Pr (I finish the paper | I go to the movies) = .25

On the preferred probability-raising view, given C and A, I have more derivative reason to go to the park than to stay home. This is so even if Pr (I finish the paper | I stay home and work on the paper) = 1.

This is in one respect similar to actualism, where the question concerns what

leave open whether there are privileged sets. In any event, for these reasons I am conflicted about my favored analysis. Positive conditional probability is a close second to the preferred analysis.

one ought to do, and the actualist answers: do that action where things turn out best. Two qualifications characteristic of actualism are (a) the value of an option (available action) is fixed by what actually happens should you choose that option, where one treats whatever happens after doing the action as an exogenous variable, even those future actions that will be available to the relevant agent, and (b) when the question is whether to do some particular action, the answer turns on whether the value of the outcome were one to do that action is greater than the value of the outcome were one to fail to do that action (go to the nearest possible world(s) to check).

The probability-raising analysis has an analogue to (a). Everything outside the set of alternatives at a context, including future actions that are not available to *S* at *C*, is treated as an exogenous variable. Though the Pr (I finish the paper | I stay home and work on the paper) = 1, and though I might have most derived reason to stay home and work on the paper when considering an *A* that includes this action string, this does not mean I have most reason to stay home. Some object to this, but the problem can be handled in the usual way. Though it is true that under the set of alternatives considered above one has more means-based reason to go to the park than to stay home, in a different context with a set of actions that includes *staying home and working on the paper*, one might well have more means-based reason to do that than to do any alternative that involves going to the park. Indeed, it is open to theorists to argue for some privileged context of choice, or a privileged set *A* in a given context, that is most appropriate for deliberation or evaluation. And the privileged *A* might include strings of actions like *staying at home and working on the paper*. Even if there is no privileged *A*, there would still be constructible As that include these strings of actions and folks who think one has most reason to stay home might have in mind not just staying home, but the string *stay home and work on the paper*.

As for (b), the probability-raising analysis on offer has no analogue. The only sets of alternatives are exhaustive and mutually incompatible ones, and we earlier rejected any baseline fixed by subjunctives that deliver pairwise comparisons, where the comparison class can shift depending on what action we are assessing. The probability-raising analysis thereby avoids some of the counterintuitive consequences of actualism, such as holding that one ought to do a terrible action when one would otherwise do something *even worse*, and despite the fact that much better options are also available. The probability-raising analysis always keeps in focus some set of available actions that is exhaustive, and so it will not lose sight of the best options available for bringing about a given end.[22]

22 Of course, the conditional probabilities for actions in the exhaustive set of alternatives will be informed by the probable future actions of the agent, which might be heinous. Worries

10. CONCLUSION

Given some conception of ends as things supported by ultimate reasons, I have tried to shed light on what it is to be a means to those ends, and how much reason is transmitted from an end to its means. The study yields a probability-raising theory, where an action is a means to an end relative to the baseline of the worst action available (probability wise). I think this is the best pure probability-raising theory, and so it is the one to compare with any alternative to the pure probability-raising approach.

APPENDIX: ANTI-ENDS AND AGGREGATION

The present theory can be extended to deal with negative reason locutions, such as "you have some reason not to ϕ," and with the problem of aggregation. Let me start with the first. It surely is intelligible to speak of reasons not to do things, as when one says, "You have a reason not to [/to not] kill innocents." Alternatively, we might speak of a reason *against* acting, as in, "You have a reason against killing innocents." These locutions raise some interesting issues. For example, when we speak of a reason for not-killing we need to know what "not" does *qua* action modifier. And when we speak of a reason *against killing* we need to know what reasons *against* action are and how they differ from reasons *for* action.

Let me focus on the first locution, and what "not" might be doing as an action modifier. We have already seen candidates for the referent of a "negated-action" clause—those considered in the discussion of baselines. For some a_i, "not-a_i" could refer to:

(1) $r(a_i)$—we assume here the nonstandard view of refraining in section 3,
(2) doing something in A other than a_i, or
(3) whatever occurs in the nearest possible world where a_i does not.

There are problems with each proposal. According to (1), "You have a reason not to [/to not] kill innocents" would refer to a reason for refraining from killing innocents. You act in accordance with this reason so long as you refrain from killing innocents. The problem here is that this refraining is typically one action

about this will be met as above, where we can restrict our attention to certain contexts of choice and sets of actions. Indicative conditionals complicate matters. If I say "If/given that I will not work on the paper here at home, I have most reason to go to the coffee shop" the antecedent is influencing the conditional probability assessments by restricting the possible worlds over which the probabilities are determined. If we calculate Pr(I finish the paper | I stay home) over worlds in which I do not write the paper at home, it might well be less than Pr(I finish the paper | I stay home) calculated over a wider set of possible worlds.

among many available ones. So suppose the relevant A = {kill innocents, refrain from doing so, buy some ice cream}. Again, we assume that refraining from killing innocents is incompatible with buying ice cream. This might seem wrong, for by having ice cream you do not kill, and you might think this is enough to thereby refrain from killing. That is right under some standard conceptions of what it is to refrain, which can be handled under option (2). However, to see all the options clearly we are working within a nonstandard conception of refraining in (1). For us, doing some action that precludes killing does not entail that one refrains from killing. Refraining from killing can be incompatible with other actions that entail that one does not kill.

Now, you have very weighty reason not to kill innocents, and some reason to buy some ice cream. Intuitively, you should buy ice cream, for that is what you have most reason to do. But if "weighty reason not to kill innocents" refers to weighty reasons *for* refraining from killing them, you would have most reason to refrain from killing innocents, for presumably the reason not to kill innocents is far weightier than your reason to buy ice cream. No ice cream for you! That is sad and wrong. For by buying ice cream you also avoid killing innocents—you can get that result without refraining from killing innocents, considered as an independent action incompatible with the other options.

It looks like (2) fixes this problem, in part by dropping the nonstandard view of refraining. According to it, "reason not to kill innocents" in our ice cream example would refer to reason *for* (refraining from killing innocents or buying ice cream). It is not entirely clear what this amounts to, for one can do either of these alternatives to killing innocents, and we want to know how much reason we have for *each* one, not for doing one or the other. One way forward is to let the strength of the reason not to kill innocents to be a reason of that strength to refrain from killing and a reason of that strength for buying ice cream. This would generate the right results for the case. You would have great reason for refraining from killing innocents and great reason for buying ice cream, plus some additional reason to buy ice cream, presumably based on its tastiness. On this view, you have most reason to buy ice cream, which is a happy result.

However, (2) cannot account for the asymmetry between tragic dilemmas and delightful dilemmas. Think of a version of Williams's Jim and the Indians case, where Jim must (a) accept the invitation and kill one person or (b) decline the invitation to kill one person, which will lead to the death of twenty by the hand of a chief. Intuitively, there are only negative things to say about the options; that is, there are reasons *against* each one. That is what makes the choice so difficult. But under (2) we are really thinking of reasons for each action—what is naturally thought of as a reason not to do one of the options, or against one

of the options, is really a reason for doing some alternative. What is naturally thought of as a reason not to decline the invitation, for instance, is really a reason to accept and kill.

By itself, this is a little odd. But it is particularly odd when we compare the case with a delightful dilemma. In a delightful dilemma, we might imagine these available actions: (a) out of a group of deserving people, pick one person to be given a genie's wish, or (b) decline to pick, in which case the genie will choose someone. If the only reasons in play are reasons to benefit others, intuitively there are only positive things to say about the options—that is, there are reasons *for* each one. That is what makes this choice difficult. What makes it delightful, and so structurally different from its tragic cousin, is that there are good reasons for each available action, rather than against each option. If (2) is correct, however, there is no such structural difference, for both cases feature reasons *for* each available action. Thus, (2) does not respect the structural difference between tragic and delightful dilemmas.

This can be put in terms of a difference in the *valence* of reasons, where a negative valence toward one action is not the same as a positive valence toward its alternatives. In other words, we should avoid interpreting these valences like vectors and their magnitudes. A vector force of negative magnitude, $-m$, in one direction is equivalent to a vector force of m in the opposite direction. By contrast, I am suggesting that a reason against an action of strength $-w$ is not equivalent to a reason of strength w for "opposing" actions.[23]

Last, (3) fails for similar reasons as (1) and (2). It has the problems of (1) insofar as it identifies a single probabilistically open alternative to receive all the weight of a reason not to a_i. There might be other probabilistically open alternatives other than the one of the nearest possible world where a_i is absent, and reasons to do these should not be outweighed by a weighty reason that favors the nearest possible world where a_i is absent. It has the problems of (2) insofar as it fails to respect a structural difference between delightful and tragic dilemmas. Through the lens of (3), all reasons can be interpreted as positive reasons for doing things or for states of affairs.

In the end, I think we should interpret the locution "reason for not a_i-ing" as referring to a reason *against* a_i-ing, where a reason against is a reason with negative weight not to be assimilated with a reason of positive weight for a_i's

23 Cf. Greenspan ("Asymmetrical Practical Reasons"), who uses a structural difference like the one defended here to think about how different reasons relate to requirements. I understand that Snedegar ("Reasons for and Reasons Against") also has a paper that discusses these issues, but I have discovered this paper only recently and have not had time to fully consider and benefit from it.

alternatives. We can get negative weights out of a probability-raising analysis of means by positing anti-ends—those things one has ultimate *negative* reason to do (/ultimate reason *against*). Anti-ends transmit their negative weight to their means in the same fashion as ends transmit positive weight to their means. In the big picture, E includes both ends and anti-ends, where for some A, each end in E transmits some positive weight to its means in A, and each anti-end in E transmits some negative weight to its means in A.

With the resources of this paper it would be nice to provide an account of aggregation into oughts. Though that must be left for another day, here are some considerations to bear in mind. First, there is the vexing case where a set of ends E has a plurality of members, and where the conditional probabilities of some ends are not probabilistically independent of one another, and particularly not independent given some means. I am thinking of cases where the Pr (e_1 and $e_2 \mid a_i$) is not calculable from Pr ($e_1 \mid a_i$) and Pr ($e_2 \mid a_i$). If the action is staying home, for example, it matters whether e_1 is finishing the paper and e_2 is grading student papers (likely not compossible), or whether e_1 is having dinner and e_2 is listening to music (likely compossible and perhaps co-probable). Even if each end is equally important and even if staying home raises the probability of each of these ends individually by .5, staying home might raise the Pr (finish paper *and* listen to music) more than the Pr (finish the paper *and* grade papers). For this reason, we might not be able to aggregate our derived reasons into oughts simply by adding up the weights transmitted from each end individually to each of its means individually.

Second, we might want aggregation to make room for incommensurability and undercutting defeat. To do so, we might not assume that each $e_i \in E$ can be assigned weight on the same scale. Some ends, you might think, have a more significant kind of weight than others, and some anti-ends transmit reasons against that should be thought of more like side constraints than weights to be traded off with other weights. One then needs to keep track of which kinds of weights get transmitted to means, where any $a_i \in A$ might be supported to different degrees by different kinds of weight. Again, this prevents a simple additive approach. But there are other options. One could lexically order the scales of weight so that one ought to do that $a_i \in A$ that has maximal weight on the highest-ranked scale in which any action registers. Or one could identify scales of negative weight that are most grievous, exclude any action with significant weight in the grievous scales and then maximize the positive weight of the remaining actions. Last, one could take a weighted average of the weights registered in all scales, giving the greatest weights to the highest-ranked scales of positive value and the most grievous scales of negative value. On this view, one can first aggregate within

each scale once the means-based reasons are identified, and create a more complicated aggregation function for weights on different scales, effectively placing them on the same scale after all.

More complicated aggregation functions are surely available to handle the most refined normative palettes. Importantly, given the modeling here, any incommensurability, defeat, enabling, or the like is to be handled either at the level of fixing ends at a context, or in the function from ultimate and transmitted reasons at a context to oughts. All this leaves open the possibility that shifts in *C* can engender shifts not only in the set of ends/anti-ends and the weights of ultimate reasons for/against them, but also shifts in how the aggregation function moves from reasons to oughts.[24]

University of British Columbia
mbedke@mail.ubc.ca

REFERENCES

Bedke, Matthew S. "The Iffiest Oughts." *Ethics* 119, no. 4 (July 2009): 672–98.

Behrends, Jeff, and Joshua DiPaolo. "Finlay and Schroeder on Promoting a Desire." *Journal of Ethics and Social Philosophy* (December 2011).

———. "Probabilistic Promotion Revisited." *Philosophical Studies* 173, no. 7 (July 2015): 1735–54.

Bratman, Michael E. "Intention, Practical Rationality, and Self-Governance." *Ethics* 119, no. 3 (April 2009): 411–43.

Coates, D. Justin. "An Actual-Sequence Theory of Promotion." *Journal of Ethics and Social Philosophy* (January 2014).

Darwall, Stephen L. *Impartial Reason*. Ithaca: Cornell University Press, 1983.

Evers, Daan. "Humean Agent-Neutral Reasons?" *Philosophical Explorations* 12, no. 1 (2009): 55–67.

Finlay, Stephen. *Confusion of Tongues*. Oxford: Oxford University Press, 2014.

———. "The Reasons that Matter." *Australasian Journal of Philosophy* 84, no. 4 (2006): 1–20.

Greenspan, Patricia. "Asymmetrical Practical Reasons." In *Experience and Analysis*, edited by Maria E. Reicher and Johan C. Marek, 387–94. Vienna: ÖBV & HPT, 2005.

24 I would like to thank Niko Kolodny, Ryan Millsap, and Jacob Stegenga for great conversations that helped this paper along. I also thank several anonymous referees for helpful comments on previous drafts.

Hitchcock, Christopher. "Do All and Only Causes Raise the Probabilities of Effects?" In *Causation and Counterfactuals*, edited by John Collins, Ned Hall, and L. A. Paul, 403–18. Cambridge, MA: MIT Press 2004.

Horty, John F. "Reasons as Defaults." *Philosophers' Imprint* 7, no. 3 (April 2007): 1–28.

Jackson, Frank, and Robert Pargetter. "Oughts, Options, and Actualism." *The Philosophical Review* 95, no. 2 (April 1986): 233–55.

Kolodny, Niko. "Instrumental Reasons." In *The Oxford Handbook of Reasons and Normativity*, edited by Daniel Star. Oxford: Oxford University Press, forthcoming.

Lin, Eden. "Simple Probabilistic Promotion." *Philosophical and Phenomenological Research*, forthcoming.

Raz, Joseph. "Instrumental Rationality: A Reprise." *Journal of Ethics and Social Philosophy* (December 2005).

———. "The Myth of Instrumental Rationality." *Journal of Ethics and Social Philosophy* 1, no. 1 (April 2005).

Schroeder, Mark. "Means-End Coherence, Stringency, and Subjective Reasons." *Philosophical Studies* 143, no. 2 (March 2009): 223–48.

———. *Slaves of the Passions*. Oxford: Oxford University Press, 2007.

Sharadin, Nathaniel. "Checking the Neighborhood: A Reply to DiPaolo and Behrends on Promotion." *Journal of Ethics and Social Philosophy* (February 2016).

———. (2015) "Problems for Pure Probabilism about Promotion (and a Disjunctive Alternative)." *Philosophical Studies* 172, no. 5 (May 2015): 1371–86.

Snedegar, Justin. "Contrastive Reasons and Promotion." *Ethics* 125, no. 1 (October 2014): 39–63.

———. "Reason Claims and Contrastivism about Reasons." *Philosophical Studies* 166, no. 2 (November 2013): 231–42.

———. "Reasons for and Reasons Against." *Philosophical Studies* (forthcoming).

Stegenga, Jacob. "Probabilizing the End." *Philosophical Studies* 165, no. 1 (August 2013): 95–112.

Suppes, Patrick. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company, 1970.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.212

ARE ALL NORMATIVE JUDGMENTS DESIRE-LIKE?

Alex Gregory

If I come to think that I ought to go to Sweden, we might think that this judgment is somewhat appetitive: if I really think this, I must be somewhat inclined to go. But in contrast, if I judge that *you* ought to go to Sweden, it is far less clear that this involves any kind of inclination on my part: I might really think that you ought to go, but need not be at all in favor of your doing so (indeed, perhaps I would much prefer you to shirk your duties and stay). Other-regarding normative judgments seem to be a matter of mere recognition, not inclination. This casts doubt on noncognitivist views according to which all normative judgments are desire-like. But it fits much better with theories in the vicinity of desire-as-belief, which identify only *some* normative judgments with desires. So I shall argue.

This paper is split into seven sections. Section 1 describes a natural way of formulating noncognitivism, which I label *conativism*. Section 2 describes the motivation argument, and presents a version that escapes some standard criticisms of that argument. In section 3, I argue that other-regarding normative judgments present a problem for the motivation argument, and indeed present a problem for conativism itself. Sections 4 and 5 consider two possible replies. Section 6 very briefly describes how the problem relates to the Frege-Geach problem. Section 7 argues that some other theories—such as desire-as-belief—may be able to accommodate the motivational role of normative judgment without falling prey to the same problem.

1. CONATIVISM

People have a variety of views about what is good, bad, right, wrong, justified, and so on. It is helpful to think of these as views about normativity (where this may include but is certainly not exhausted by, moral normativity).[1] I follow tra-

1 Throughout this paper, I wholly ignore epistemic normativity, which raises too many issues to be adequately discussed here. To the extent that a conativist analysis of epistemic norma-

dition and stipulatively use the word "judgment" to refer to the state of mind (whatever it is) that such views consist in. With this terminology, we can formulate a theory:

Conativism: All normative judgments are desires.

We can think of conativism as one particular kind of noncognitivist theory. Noncognitivists deny that normative judgments are beliefs. Conativism adds to noncognitivism by also making a claim about what normative judgments are: desires. I take it that conativism represents a central strand of the noncognitivist tradition. For example, one classic argument for noncognitivism is the motivation argument, which appeals to the fact that normative judgments motivate us in a way that only desires can.[2] If this moves us to accept noncognitivism, it should move us to accept the conativist kind of noncognitivism, since it establishes the conclusion that normative judgments are desires, not merely the conclusion that they are not beliefs.

In this paper, I object to conativism. I thereby leave open that there might be other noncognitivist views that are plausible and that escape the objection I present against conativism. For example, I shall not discuss noncognitivist views that abandon the motivation argument entirely and treat all normative judgments as states of mind that are neither beliefs nor desires. I shall also not discuss noncognitivist views that claim that whereas some normative judgments are desires, others are some other noncognitive state of mind. I tend to think that conativism captures an important strand of the noncognitivist tradition, and that noncognitivist views other than conativism are likely to lose some of the advantages that noncognitivism is supposed to have over cognitivism. But other than in a brief note, I shall not address such issues.[3] From here onward

tive judgments is implausible, that would further support my general conclusion that not all normative judgments are desires.

2 See, e.g., Blackburn, *Spreading the Word*, 187–89; as well as Smith, *The Moral Problem*.

3 First, some noncognitivists might claim that normative judgments are states of *approval* and *disapproval*. But if they also wish to maintain that normative judgments can motivate, it seems as though such noncognitivists will need to *object* to the Humean view that only desires can motivate us! And once they do that, it is far less clear that there is any real reason to deny that normative judgments are beliefs to begin with. For this reason, it seems more likely that such noncognitivists are tacitly thinking of states of approval and disapproval as desire-like in the relevant ways, and so are really conativists in disguise. Second, some noncognitivists might claim that some normative judgments are desires, but that other such judgments are (non-desire-like) states of approval and disapproval. But if different normative judgments are not even the same state of mind as one another, such noncognitivists will have an even harder time with another classic problem for noncognitivism: Making sense of normative disagreement and inconsistency. It is not at all clear how one mental state

my focus is simply on conativism: the reader may decide for themselves how my discussion bears on noncognitivism more broadly.

Despite this restriction of focus, my objection to conativism will extend to nearby views in three ways. First, conativism is a view about the nature of a mental state: normative judgment. By itself, it says nothing about normative language. Some noncognitivists defend expressivism, which combines noncognitivism about normative judgment with a further claim about normative language: expressivists say that the meaning of normative utterances is determined by the state of mind they express. In what follows I continue to focus on the nature of various states of mind rather than the meanings of utterances, but my objection might nonetheless have implications for expressivism insofar as many expressivists incorporate conativism into their view.

Second, quasi-realists are in the noncognitivist tradition but nonetheless claim that normative judgments are beliefs. Insofar as such views are coherent, they reconcile these claims by distinguishing two different kinds of belief: full-blown beliefs and states that are beliefs only in some minimalist sense of "belief."[4] They then claim that normative judgments are beliefs only in the second minimalist sense.[5] Nothing stops conativists from adopting quasi-realism: they merely need to claim that normative judgments are both desires and beliefs, but the latter only in some minimalist sense. My objection to conativism applies equally to quasi-realist conativism.

Third, some noncognitivists claim that normative judgments are not desires. But if these authors nonetheless maintain that normative judgments have the same motivational profile as desires, that suffices for my purposes.[6] When I object to conativism below, it is this aspect of it that I focus on, and it is therefore unimportant whether some noncognitivist denies that normative judgments are desires for reasons that are independent of their motivational profile. One way of putting the point is to say that when I object to conativism, I really object to the claim that all normative judgments have the desire-like direction of fit.[7] To that extent, my objection applies to any noncognitivist view that endorses this claim, even if they deny that normative judgments are desires, strictly speaking.[8]

---

with one content could be inconsistent with a mental state of a distinct kind with a distinct content.

4 See, e.g., Dreier, "Meta-Ethics and the Problem of Creeping Minimalism."

5 See again Dreier, "Meta-Ethics and the Problem of Creeping Minimalism," especially 26–29.

6 For this combination of claims, see Blackburn, *Ruling Passions*, especially 9–10, 13–14, 66; and, plausibly, Gibbard, *Wise Choices, Apt Feelings*, 55–56, 75)

7 For this notion, see, e.g., Humberstone, "Direction of Fit"; Smith, *The Moral Problem*, 115

8 Do my arguments also extend to "hybrid" noncognitivist theories, which analyze normative

In short, here I focus on conativism, but my arguments seem likely to extend to many theories in the noncognitivist tradition, either because they incorporate conativism or because they incorporate claims that are in the relevant respects close enough to conativism. But I leave open that there might be some better noncognitivist theories that do not commit to conativism or anything relevantly similar. In that case, my arguments against conativism at least highlight some constraints that any plausible formulation of noncognitivism must meet.

2. THE MOTIVATION ARGUMENT

One central argument for conativism is the motivation argument.[9] The basic idea behind the motivation argument is that because normative judgments bear a special connection to motivation they must be desires. This argument is one of the main weapons that conativists have against rivals, such as cognitivist naturalists. But formulating this argument in a manner that is both precise and plausible has proved difficult. Here I suggest that it is attractive to formulate the argument in roughly the following manner:

P1: Normative judgments can motivate.
P2: Only desires can motivate.
So, C: Normative judgments are desires.[10]

This argument is still somewhat ambiguous, in that the scope of P1 and C is unclear. In later sections, I shall argue that when we disambiguate the scope of these premises, the argument is either unsound or else fails to support conativism. But before we get to that, I want to first note the virtues of formulating the motivation argument along the above broad lines: the relevant points will not hinge on how we disambiguate the scope of P1 and C. Unlike other formulations of the motivation argument, this formulation focuses on the capacities (powers) of the relevant states of mind.[11] In this respect, it is parallel to the argument that

judgments as combinations of desires and beliefs (see, e.g., Fletcher and Ridge, *Having It Both Ways*; Ridge, *Impassioned Belief*)? So far as I can see, there is no straightforward answer to this question, and it may depend on the hybrid theory in question. But it is worth noting that at least one prominent hybrid noncognitivist seems to endorse the view that I describe and criticize in section 5 (Ridge, *Impassioned Belief*, 19, 177–78).

9 See, e.g., Blackburn, *Spreading the Word*, 187–89, and *Ruling Passions*, 70; Gibbard, *Thinking How to Live*, 11–13; Hare, *The Language of Morals*, 1; Stevenson, "The Emotive Meaning of Ethical Terms," 16.

10 Cf. Snare, *Morals, Motivation and Convention*, 58.

11 With this in mind, I am treating "can" as a predicate ascribing a power, though a modal version of the argument could also be formulated.

because H_2O can quench thirst, and only water can quench thirst, water is H_2O. In this kind of argument, note the obvious fact that the premises by no means imply that the relevant powers are always being exercised. This formulation of the motivation argument relies on the claim that normative judgments have the power to motivate, but is consistent with the fact that they often fail to exercise that power, such as when we are weak willed, or make other contrary normative judgments. We will see the significance of this shortly when I contrast this argument with other formulations of the motivation argument.

I take it that P1 of this argument—the claim that normative judgments can motivate—is attractive.[12] It is attractive to think that people can do things because they think they have good reason to do them, and, to repeat, the premise permits that people can be left cold by their normative judgments. Once that is acknowledged, it is hard to see any opposition to the premise that is not theory driven. P2 claims that only desires have the capacity to motivate us to act. This is the Humean theory of motivation, in one form. Again, I take it that this claim is attractive. When we explain people's motivations, it seems as though any explanation must ultimately appeal to their desires.

These brief remarks are not intended to demonstrate P1 and P2 above. I like to think that there is at least a good *prima facie* case for accepting them, and in turn a *prima facie* case against views that are inconsistent with the conclusion of the above argument. But adjudicating this dispute is not my main concern in this paper. Rather, I want to address whether these premises can have their scope disambiguated in a manner that allows them to support conativism. So in what follows I shall simply assume that P1 and P2 above are *broadly* plausible, and shall instead focus our attention just on the appropriate scope of P1 and C.

For the purposes of this paper, I treat the Humean theory as the positive claim above, and not as the negative claim that beliefs do not have the capacity to motivate us to act. That negative claim would fit less neatly with quasi-realist conativist views on which normative judgments are beliefs in addition to their being desires. Further, that negative claim is not well supported by the best arguments for the Humean theory. One well-known argument for the Humean theory is Smith's, which appeals to the distinctive direction of fit of desire.[13] If successful, that argument shows that only states with the desire-like direction of fit can motivate, but does not directly show that any state with a belief-like direction of fit cannot motivate, since it leaves open that some states have both di-

12 Cf. Broome, "Reasons and Motivation," 139; Dancy, *Moral Reasons*, 22–23; Ridge, *Impassioned Belief*, 50

13 Smith, *The Moral Problem*, 116.

rections of fit at once.[14] But the best argument for the Humean theory is simply that such a view is extremely parsimonious and yet has great explanatory power. It promises to explain all human behavior by appeal to just two kinds of mental state: desires and means-ends beliefs. That argument obviously appeals to the explanatory power of the claim that only desires can motivate, not to the lack of explanatory power of a theory on which beliefs can motivate, and so supports the positive claim P2 and not the negative claim that I ignore.

Relatedly, we might worry that P2 ought to be formulated as, "Only pairs of desires and means-ends beliefs can motivate."[15] If we formulated it that way, the above argument would be invalid: at best we could conclude that normative judgments are either desires or means-end beliefs. I am not sure whether sympathizers of the Humean theory should be happy with this reformulation of the premise: we might think that desires are the real motivational workers, and that our means-ends beliefs do not themselves partly motivate us, but instead just channel the motivational powers of desire in new directions. But regardless, even if we reformulated the premise in this manner, we could easily reformulate P1 of the argument to maintain the validity of the argument. We could just reformulate P1 to say that normative judgments can motivate, and can do so in a way that means-ends beliefs cannot. After all, the thought driving P1 is that there is some special connection between normative judgment and motivation, and that thought is lost if we permit that normative judgments merely play a role in motivation that ordinary beliefs can also play. For ease, in what follows I stick with the simpler formulation of the argument above.

Let me briefly highlight one virtue of my formulation of the motivation argument by contrasting it with two others. First, a standard way of thinking of the motivation argument has it appeal to the following premise instead of P1 above:

P1*: The judgment that I ought to ϕ necessarily motivates.

P1* is sometimes called "classic" or "mad dog" judgment internalism.[16] But as many have noted, P1* is implausible.[17] Through weakness of will, we might fail to be at all moved by judgments about what we ought to do. (Here I mean not merely that we might fail to act on some judgment, but that we might fail to have

14 Of course, Smith does also argue that no state can have both directions of fit at once (*The Moral Problem*, 117–25). But that is a further independent claim, and a dubious one at that. See Little, "Virtue as Knowledge," 63–64; Price, "Defending Desire-as-Belief," 120–21.

15 See, e.g., Smith, *The Moral Problem*, 92; Davidson, "Actions, Reasons, and Causes," 3–4.

16 See, e.g., Björklund et al., "Recent Work on Motivational Internalism," 125; Gibbard, *Thinking How to Live*, 153; Hare, *The Language of Morals*, 163–69.

17 See, e.g., Svavarsdóttir, "Moral Cognitivism and Motivation," 176–83.

any motivation to act whatsoever, even a motivation that is outweighed.) Since P1* is implausible, it might seem that the motivation argument for conativism cannot be sound. But P1, above, unlike P1*, permits the possibility of weakness of will, and thus evades this objection.

Since P1* is implausible, others have formulated the relevant claim in other, more modest ways. One popular alternative says something like the following:

> P1**: The presence of the judgment that I ought to ϕ normally/rationally entails the presence of motivation to ϕ.[18]

P1** has a definite advantage over P1*, since it permits weakness of will. Perhaps some claim along the lines of P1** is true. But this comes as a hollow victory for fans of the motivation argument, since it is hard to formulate the motivation argument in a manner that can appeal to this premise and validly get to the conclusion that normative judgments must be desires.[19] Since P1** is a claim about mere covariation, it is consistent with the thought that normative judgments influence motivation only through their influence on independent desires, and this is consistent with denying conativism.[20] P1 is more significant since it makes a claim not about mere covariation, but instead more directly about explanation.[21]

In short, while P1* makes the motivation argument unsound, P1** is likely to make it invalid. In contrast, P1 of the argument as I have formulated it seems modest enough to be plausible but yet bold enough to make the argument valid. To repeat, my goal here is not to show that the argument as I formulate it is conclusive, but rather to show that it has *prima facie* appeal and can survive standard objections. This is enough to set the scene for the rest of the paper, in which I examine what this argument might show.

18 See, e.g., Korsgaard, "Skepticism about Practical Reason," 15; Smith, *The Moral Problem*, 12, 61; van Roojen, "Moral Rationalism and Rational Amoralism," 499.

19 Brink, "Moral Motivation," 7–8; Svavarsdóttir, "Moral Cognitivism and Motivation," 165–66n.

20 Smith, *The Moral Problem*.

21 We could of course modify P1** so that it too is a claim about explanation: we might say, for example, that rational agents are motivated by their normative judgments. Such claims might well be plausible, and could also play a role in the motivation argument. But in that form the relevant claims add nothing to the motivation argument beyond what P1 already provides. As such, the proposed modified version of P1** is unnecessarily bold, requiring us to defend claims (e.g., about rationality) that go beyond the commitments of P1 but that do nothing to make the motivation argument more forceful.

3. THE PROBLEM: OTHER-REGARDING NORMATIVE JUDGMENTS

We should ask a simple question about the motivation argument: whether P1 and C are supposed to be read as being universally quantified or not. Since conativism is the view that all normative judgments are desires, conativists should presumably formulate the argument with both claims universally quantified. In that form, let us call it the bold motivation argument. It reads:

Universal-P1: All normative judgments can motivate.
P2: Only desires can motivate.
So, Conativism: All normative judgments are desires.

In this section, I first offer some counterexamples to Universal-P1. This suggests that the bold motivation argument fails to establish conativism. The bold motivation argument seeks to establish conativism by appealing to the motivational powers of our normative judgments, but some of our normative judgments do not have any motivational powers. After presenting this argument against the bold motivation argument, I suggest that the very same counterexamples cast doubt on conativism itself.

It seems as though other-regarding normative judgments have no motivational powers of their own. For example, if Jane judges that Jeff ought to buy his child a birthday present, it seems that this judgment has no power to motivate Jane to do anything. It is a judgment about what Jeff should be doing, and by itself has no bearing at all on what Jane herself will do. When Jane judges that Jeff ought to ϕ, that is a matter of recognition, not inclination. This casts doubt on Universal-P1.

This point can be obscured by the fact that other-regarding normative judgments can play a role in inference to further normative judgments that do have motivational powers. For example, if Jane *also* judges that she ought to assist others in doing their duty, she might infer that she ought to help Jeff buy the present, and this judgment might motivate her. But here the motivational power is infused only by the addition of a further, self-regarding normative judgment: without it, the original other-regarding normative judgment is motivationally inert. So this possibility fails to show that other-regarding normative judgments have motivational powers of their own.

Similar reasoning undermines other possible reasons for endorsing Universal-P1. For example, one might think that if Jane judges that Jeff ought to keep his promises, but finds that he does not, this might motivate her to avoid him. Or one might think that if Jane judges that Jeff does not invest his money as he ought, this might motivate her to avoid lending him money. But plausibly what

really motivates Jane in the first case is the judgment that *she* ought not trust people who do not keep their promises, and what really motivates Jane in the second case is the judgment that *she* ought not lend her money to people who are bad with money. As such it is again doubtful that her other-regarding normative judgments themselves have motivational powers.

The starkest counterexamples to Universal-P1 are those in which one person judges that another ought to do something that conflicts with the goals of the first. Most extremely, imagine that Jane is a consistent egoist, who judges that everyone ought only to promote their own well-being. Jane might thereby judge that Jeff, in his dealings with her, ought to use and abuse her. This judgment, it seems clear, would have no motivational power over Jane. The point can be seen equally well in less extreme cases: in a prisoner's dilemma, Jane might judge that her partner ought to rat, but it is doubtful that this alone can motivate her to do anything. Or for a final example, I might judge that you ought to save your mother at the expense of mine, but not want you to do so. In general, we find stark counterexamples to Universal-P1 whenever the relevant norms are believed to be agent-relative, as many prudential and moral norms are believed to be. In such cases, an agent can think that some norm applies to another but not to themselves, and as such be left cold by their recognition of that norm.

One might reply that Universal-P1 says only that normative judgments *can* motivate, and for that reason it is consistent with the fact that they often fail to do so. But the worry is that in cases like those above it is not even plausible that the relevant judgments *could* motivate us: there are *no* conditions under which other-regarding normative judgments motivate. It just does not seem intelligible for someone to act in some way because they recognize that *someone else* ought to do something. Certainly, agents in the cases above are not merely being akratic: it is not as though egoists, or prisoners in the prisoner's dilemma, are being weak willed when they refuse to help others do what they ought to do. Indeed, if we think that there is a rational requirement not to be akratic, we would thereby commit ourselves to the claim that (e.g.) the prisoner who fails to persuade her partner to rat is being *irrational*, and this is highly implausible. Nothing need be irrational or even abnormal about an agent who judges that *they* have no reason to comply with a norm that they judge governs someone else but not themselves.

Another way of looking at this problem is to remember that the first premise of the motivation argument is a form of judgment internalism. But no plausible formulation of that view makes a claim about all normative judgments. Judgment internalism, as standardly formulated, makes a claim only about self-regarding normative judgments (see references above). So the motivation argument, if it is supposed to support conativism, must appeal to something bolder

than judgment internalism: the view that my normative judgments about anyone have motivational power over me. This claim is far from obvious.[22]

In light of this, it seems reasonable to deny that other-regarding normative judgments have motivational powers, and in turn reasonable to deny Universal-P1 of the bold motivation argument. To that extent the bold motivation argument is unsound and so fails to establish conativism. Perhaps there is some other way of formulating this argument, but conativists would have to show what that formulation is, and show that it avoids commitment to Universal-P1 above. Certainly, we should not assume that there is any straightforward argument from judgment internalism and the Humean theory of motivation to conativism, since judgment internalism is best formulated as a claim about only some specific normative judgments, and need not teach us anything about the nature of the whole class.

So far I have suggested that the motivation argument needs to be reformulated in some nonobvious way if it is to soundly support conativism. Perhaps conativists might simply drop the motivation argument and maintain their view by appeal to other arguments, though we might worry that this amounts to abandoning one of the most persuasive arguments for their view. But there is a further and larger problem: the counterexample to Universal-P1 seems to extend to cast doubt on conativism itself.

If other-regarding normative judgments are motivationally inert, that strongly suggests that they are not desires. Indeed, this straightforwardly follows if we accept the popular theory that analyzes desires precisely in terms of their capacity to motivate.[23] It is true that desires are normally thought to motivate only when combined with suitable means-ends beliefs (see section 2 , above), but this does nothing to help the conativist: it is clear that we do often have the relevant means-ends beliefs and still lack the corresponding motivations. For example, Jane might judge that Jeff ought to use and abuse her, and believe that she can get him to do so by anonymously sending him the works of Ayn Rand, but still not have any motivation to do so. It seems to follow that her judgment that Jeff ought to use and abuse her is not a desire that he use and abuse her, and it is not clear what other desire it might be.

Perhaps we could add other conditions that are necessary for desires to motivate, or else distinguish between motivating and non-motivating desires.[24] Such claims might allow us to say that other-regarding normative judgments might be desires even if they are motivationally inert. But even if this line of reasoning

22 Cf. Gibbard, *Wise Choices, Apt Feelings*, 100–1; Ridge, *Impassioned Belief*, 19.

23 See, e.g., Smith, *The Moral Problem*, 92–129; Stalnaker, *Inquiry*, 15.

24 See, perhaps, Mele, "Motivation."

could be maintained, it is nonetheless independently implausible that there is a necessary connection between other-regarding normative judgments and desire. If Jane judges that Jeff ought to buy his child a birthday present, that seems to leave completely open whether she altruistically hopes he does, spitefully hopes he does not, or just fails to care either way about what she sees as Jeff's business. And again, more starkly, a committed egoist surely need not desire that others do what they ought to do, and someone in a prisoner's dilemma can recognize that their partner ought to rat without desiring that they do so. Even independently of anything we say about motivation, it is not plausible that we necessarily desire that others do what they ought.

In short, when we reflect on other-regarding normative judgments, it seems as though they serve as counterexamples to the claim that all normative judgments have motivational powers, and as counterexamples to the claim that all normative judgments are desires. Reflection on such judgments seems to thereby cast doubt on the motivation argument for conativism, and on conativism itself. When we make judgments about what others ought to be doing, that seems to be a matter of mere recognition, and need not involve any inclination on our own part. In the following two sections, I consider two possible replies open to the conativist. In each case, I argue that the relevant response opens the resulting theory to other objections. We should conclude that the objection above places a significant constraint on any plausible formulation of conativism, and that it is at least unclear whether any independently plausible conativist theory can meet that constraint.

4. Reply 1: Reactive Attitudes

We might think that certain reactive attitudes—such as blame and guilt—are important aspects of morality.[25] This might encourage the conativist to claim that other-regarding normative judgments are desire-like after all. In this section, I focus on our dispositions to *blame* others for wrongdoing, though I take it that the relevant arguments extend in obvious ways to nearby alternatives such as our dispositions to shun or to punish wrongdoing, or to praise or reward virtue.

The conativist might press the above line of reasoning above in two different ways. First, they might take the fact that we are disposed to blame others for wrongdoing as evidence that we desire others to act rightly. Second, they might identify our disposition to blame others for wrongdoing with a desire: they might claim that other-regarding normative judgments just *are* (or involve) desires to blame others under relevant circumstances.

25 See, e.g., Blackburn, *Ruling Passions*, 8–14; Gibbard, *Wise Choices, Apt Feelings*, 41–45.

The first of these views is still committed to the counterintuitive claim that people necessarily want others to do what they ought to do. It tries to justify this claim by appeal to our emotional dispositions, but it seems that any such justification will be defeated by the fact that it is independently implausible that people *do* necessarily want others to do what they ought to do. That is just what I argued above: it is doubtful that Jane necessarily wants Jeff to buy his child a present, and very implausible that egoists, or people in the prisoner's dilemma, necessarily want others to do what they ought. Even if many of us often do want others to do what they ought, this connection seems highly contingent, and to that extent, we cannot identify other-regarding normative judgments with desires for others to do the relevant things.

But the second of the above views is little better: it is equally implausible that people necessarily want to blame others for wrongdoing. One simple worry is that someone might judge an act to be wrong but blameless: an agent might endorse an other-regarding normative judgment but have no corresponding desire to blame them.[26] But even if we set this aside, it is clear that this second view seems plausible only if we focus solely on moral normativity. It might be true that there is some necessary connection between our moral attitudes and our dispositions to blame. But to the extent that this is plausible, it seems plausible because we are thinking of morality as a kind of social phenomenon.[27] To that extent, other normative domains that are less social in nature seem to lack any necessary connection with our reactive attitudes. Think, for example, about prudential judgments. Imagine that Jane judges that prudence requires Jeff to buy new running trainers. It is doubtful that she must thereby want him to buy those trainers: she might make this judgment and yet not really care whether he does what he ought. Given that she might not care whether Jeff does what he ought, it seems no huge step to suppose that she might also fail to care what happens to Jeff as a result of his failing to do what he ought.[28] If she does not care much about Jeff at all, she might care neither whether he is prudent nor whether he is chastised when he is not.

Note that I need not claim that such attitudes are common. The point is simply that it is not a necessary condition on judging that someone else ought to ϕ that you desire to blame them if they do not. If we find some alien culture in which people have no concept of blame, it is far from clear that this would conclusively demonstrate that that same culture has no normative concepts at all.

26 Cf. D'Arms and Jacobsen, "Expressivism, Morality, and the Emotions"; Ridge, *Impassioned Belief*, 142–43.

27 Cf. Gibbard, *Wise Choices, Apt Feelings*.

28 Again, cf. Ridge, *Impassioned Belief*, 143.

A utilitarian culture is surely possible, within which people have various views about what people ought to do, but no corresponding inclinations to blame anyone for wrongdoing, since they consider it counterproductive. But this would not be possible if other-regarding normative judgments just *are* desires to blame others.

In short, it may be plausible that there is *some* connection between some normative judgments—especially moral judgments—and our reactive attitudes, such as our dispositions to blame others. But it is not plausible that there is a necessary connection between other-regarding normative judgments and our dispositions to blame others, and this casts doubt on this conativist strategy.

5. Reply 2: Gibbard

We might think that other-regarding normative judgments are not directly motivating, but that they nonetheless qualify as desires (or at least desire-like) because they have indirect motivational influence. One initial problem with this suggestion is that many states of mind have some indirect motivational influence. For example, regular beliefs have some influence on what we are motivated to do. So to maintain this suggestion, we need to find some distinctive kind of indirect motivational influence that is had by other-regarding normative judgments but not by other states of mind that the conativist wants to exclude from their theory.

So far as I can see, the best option for the conativist is to claim that other-regarding normative judgments influence motivation in a way that depends only on combining them with prior beliefs. That kind of motivational influence is not had by other states of mind, such as regular beliefs. And if we combine this thought with the common thought that any state of mind that can motivate when combined only with belief(s) is by definition a desire, we can infer that other-regarding normative judgments are desires.[29]

If Jane's judging that Jeff ought to ϕ can motivate her by being combined only with a belief about how to get Jeff to ϕ, then such a judgment just is the desire that Jeff ϕ. That is the suggestion that we have already rejected: it seems plausible that we can judge that others ought to do things we do not want them to do. The present suggestion is instead better developed in the manner explained by Allan Gibbard.[30] On Gibbard's view, other-regarding normative judgments are desires about what to do if you were in the other person's place. That is, Gibbard claims

29 See, e.g., Smith, *The Moral Problem*, 92–129; Stalnaker, *Inquiry*, 15.

30 Gibbard, *Thinking How to Live*, 49–53; see also Gibbard, *Meaning and Normativity*, 174–77; Ridge, *Impassioned Belief*, 19, 177–78. Gibbard treats normative judgments as plans rather than desires, but, as I said in section 1, this small kind of difference seems unimportant for

that Jane's judgment that Jeff ought to ϕ is a conditional desire to ϕ if she were in Jeff's place. Such a view rightly permits that Jane might judge that Jeff ought to ϕ but not want him to, and yet nonetheless treats such judgments as desires, as conativists must. And such conditional desires do indeed have only indirect motivational potential, as promised: conditional desires need to be combined with the belief that the relevant condition is met if they are to motivate. As applied in our case, we get the conclusion that Jane's judgment can motivate her, but only by being combined with the belief that she is Jeff.

We should distinguish two more precise ways we might develop Gibbard's proposal. First, the simpler option:

> Gibbard-simple: When *A* judges that *B* ought to ϕ, that consists in *A*'s conditionally desiring to ϕ if they, *A*, were in *B*'s circumstances.[31]

Second, the more sophisticated option:

> Gibbard-sophisticated: When *A* judges that *B* ought to ϕ, that consists in *A*'s conditionally desiring to ϕ if they were *B* and in *B*'s circumstances.

The difference between the proposals is that the sophisticated proposal, unlike the simple, understands other-regarding normative judgments to be desires regarding circumstances in which we have different haecceities than we in fact have. So far as I can tell, Gibbard himself endorses the second option, but for completeness I object to both, in turn.[32]

So first, there is Gibbard-simple. The first thing to note is that Gibbard-simple is counterintuitive. It is counterintuitive to say that when Jane makes a normative judgment about Jeff, she is forming a conditional desire for the eventuality that she end up in his circumstances. The view is all the more surprising once we remember that "circumstances" here has to include not only Jeff's external environment, but also anything that might be relevant to what he ought to do:

---

our purposes, since Gibbard nonetheless treats plans as being like desires in the way that they motivate. For simplicity, I continue to talk in terms of desire.

31 This sort of view may also have been held by Hare. See, e.g., *Moral Thinking*, especially ch. 7. But matters are not so clear because Hare often talks about what a person is *committed to*, and this more often sounds like a normative claim rather than a descriptive one. Gibbard-simple says that other-regarding normative judgments literally *are* desires of the relevant sort, and Hare may have meant to commit only to the weaker thesis that other-regarding normative judgments *rationally require* desires of the relevant sort. I say nothing here against this latter thesis, which might well be true, but fails to provide the conativist with what they need. See also the discussion of supervenience, below.

32 See Gibbard, *Thinking How to Live*, 50–51.

his ignorance, his character traits, his emotions, and so on.[33] So the suggestion had better be that, when Jane judges that Jeff ought to ϕ, this is a desire of hers to ϕ if she had all of his properties. And that is a conditional desire for a possibility that is extremely unlikely to occur, if it is possible at all.[34] Other-regarding normative judgments are a familiar everyday mental state, and it would be surprising if they turned out to be an attitude directed toward such bizarre counterfactual circumstances; other-regarding normative judgments certainly do not *seem* to be attitudes toward extremely remote possible worlds, but instead attitudes directed toward other people in the actual world.

This point is especially clear if we think about how children learn to make normative judgments about others: it is highly doubtful that they do so by thinking about various highly remote possibilities. If a young boy thinks that girls are not supposed to have short hair, they do not come to that thought by imagining themselves as a girl (indeed, a failure to do so is presumably part of the problem). Perhaps it might be conceded that children *sometimes* form views about others by considering the relevant counterfactuals, but it is highly doubtful that they *always* do so, especially when the relevant counterfactual possibilities are (or are seen to be) very remote. So the first concern is that Gibbard-simple is highly counterintuitive, and to that extent it seems *ad hoc*.

Gibbard-simple also faces a second objection. Gibbard-simple analyzes Jane's judgment [that Jeff ought to ϕ] as Jane's desire [to ϕ if she were in Jeff's circumstances].[35] Presumably, Gibbard-simple would also have us analyze Jane's judgment [that she ought to ϕ if she were in Jeff's circumstances] as Jane's desire to ϕ if she were in Jeff's circumstances]. Since these two judgments receive the same analysis, Gibbard-simple forces us to identify them. That is, according to Gibbard-simple, there is no difference between Jane's judgment [that Jeff ought to ϕ] and Jane's judgment [that she ought to ϕ if she were in Jeff's circumstances]. But the problem is that these two judgments are distinct. This is clearest when we consider the obvious fact that Jane might have concluded that Jeff ought to ϕ without having extended the conclusion to her own case. The reverse is also possible: Jane might have made a plan for herself in Jeff's circumstances without having actually considered what Jeff himself ought to do. By analyzing other-re-

33 Cf. Gibbard, *Thinking How to Live*, 50–51, and *Meaning and Normativity*, 174–75.

34 In passing, Gibbard claims that even if it is sometimes metaphysically impossible for one person to be in another's exact circumstances, it is nonetheless *epistemically* possible (*Meaning and Normativity*, 177). But even if this is right, it is unclear how this is supposed to dispel the oddness of the view given how epistemically remote those possibilities often are.

35 Here, and later, I sometimes use square brackets to mark the contents of attitudes.

garding normative judgments as conditional desires, we lose the ability to give an independent analysis of genuinely counterfactual normative judgments.

Might Gibbard reply that other-regarding normative judgments and genuinely counterfactual normative judgments are in fact the same state of mind, and claim that this truth is merely non-obvious to us? But that reply suggests that there is no such thing as an inference between these states of mind, and that is implausible. Plausibly, we can sometimes get people to change their minds about how they judge others precisely by having them consider the relevant counterfactual claims about themselves, and vice versa, and this would not be possible if the relevant states of mind were literally identical.

Gibbard-simple gains illusory plausibility here because normative truths supervene on nonnormative truths (henceforth: Supervenience). With Supervenience in mind, it is tempting to think that if Jeff ought to ϕ, then Jane ought to ϕ if she were in Jeff's exact circumstances. But even if Supervenience is true, that does not tell us much about Jane's judgments. Jane might fail to accept Supervenience, or more likely, might fail to accept every single implication of that truth. So Supervenience does not show that Jane's judging that Jeff ought to ϕ is the very same thing as Jane's judging that she ought to ϕ if she were in Jeff's circumstances.

Some noncognitivists have claimed that Supervenience is not a metaphysical truth, but instead a conceptual one.[36] But even this will not help. Even if Jane's judgments embody failures to accept conceptual truths, this is no bar to her having those judgments.[37] Jane might be conceptually confused and deny Supervenience. Or again, more likely, she might accept Supervenience but fail to accept various truths that are entailed by combining Supervenience with other beliefs that she holds—she might fail to combine her very abstract commitment to Supervenience with her judgment that Jeff ought to ϕ (and it is surely possible to fail to endorse every implication of a conceptual truth one recognizes).

In short, Gibbard-simple is counterintuitive, and moreover it collapses the distinction between Jane's judging that Jeff ought to ϕ and Jane's judging that she ought to ϕ if she were in Jeff's position. Those judgments are distinct, and Supervenience does not show otherwise.

I now turn to Gibbard's own preferred view, Gibbard-sophisticated. To remind you, it says:

36 See e.g., Blackburn, *Essays in Quasi-Realism*.

37 Cf. Williamson, *The Philosophy of Philosophy*, 73–133. At one point Blackburn seems to assert the reverse (*Essays in Quasi-Realism*, 122) but he gives no argument for this bold claim, and it is not required for the Supervenience argument that he presents against moral realism.

> Gibbard-sophisticated: When *A* judges that *B* ought to ϕ, that consists in *A*'s desiring to ϕ if they were *B* and in *B*'s circumstances.

It is not obvious that Gibbard-sophisticated really improves on Gibbard-simple. It does allow us to distinguish Jane's judgment [that Jeff ought to ϕ] and Jane's judgment [that she ought to ϕ if she were in Jeff's circumstances]. Gibbard-sophisticated entails that the former but not the latter consists in a desire that is conditional on Jane's being Jeff. But this is not obviously progress, since we might now worry that Gibbard-sophisticated fails to distinguish a different pair of judgments: Jane's judgment [that Jeff ought to ϕ] and Jane's judgment [that she ought to ϕ if she were Jeff and in Jeff's circumstances]. Gibbard-sophisticated will presumably give these two judgments the same analysis, and this might seem mistaken for just the same reasons as those given above. Again, by analyzing other-regarding normative judgments as conditional desires, we lose the ability to give an independent analysis of genuinely counterfactual normative judgments.

That said, here it is admittedly *somewhat* less clear-cut that these two judgments really are distinct. But whatever plausibility Gibbard-sophisticated gains here, it loses with respect to the first worry above. To whatever extent Gibbard-simple was counterintuitive, Gibbard-sophisticated is worse. Gibbard-sophisticated says that people who make judgments about what other people ought to be doing are forming desires for circumstances in which their identity differs. It is highly counterintuitive to suppose that we have conditional desires for such impossible circumstances: even if such conditional desires are possible, it does not seem that they are commonly occurring parts of our mental lives. And again, it is deeply implausible that we learn to make other-regarding normative judgments by learning to think about how our own identity and circumstances might differ: it is not clear that the average child is even capable of thinking about such matters. These implications of Gibbard-sophisticated should make us very wary of accepting it unless there is no other option.

I conclude that Gibbard's strategy is at best ad hoc and counterintuitive, and at worst inconsistent with clear distinctions between different normative judgments. I also argued above that conativists should not try to rescue their view by appeal to the connection between normative judgments and reactive attitudes. With no other obvious option on the table, I conclude that the objection stands and conativism is implausible.

6. THE FREGE-GEACH OBJECTION

It might be worth very briefly comparing the objection I have raised against conativism with the Frege-Geach objection to noncognitivism.[38] As applied to conativism, the Frege-Geach objection is that normative judgments with logically complex contents cannot be analyzed as desires.[39] We can illustrate this claim by appeal to normative judgments with negated contents.[40] Jane's judgment that it is not the case that she ought to drink tea seems hard to analyze as any desire. It is not a desire to drink tea, nor a desire not to drink tea, nor a failure to desire to drink tea. Worse, even if we could find some way to analyze this judgment as one of these desires, that would only move the problem elsewhere: the judgment more naturally associated with the relevant desire would itself now lack a suitable analysis. That is, there is no way to reduce the following four states of mind on the left in terms of the three on the right:

Judgment that she ought to ϕ ($\mathrm{J}\mathrm{O}\phi$)	Desire to ϕ ($\mathrm{D}\phi$)
Judgment that she ought not ϕ ($\mathrm{J}\mathrm{O}\neg\phi$)	Desire not to ϕ ($\mathrm{D}\neg\phi$)
Judgment that it is not the case that she ought to ϕ ($\mathrm{J}\neg\mathrm{O}\phi$)	???
Failure to judge that she ought to ϕ ($\neg\mathrm{J}\mathrm{O}\phi$)	Failure to desire to ϕ ($\neg\mathrm{D}\phi$)

38 See Schroeder, *Being For*, for comprehensive discussion.

39 Often, the Frege-Geach objection is expressed as the problem for expressivism of accounting for the meaning of logically complex sentences that employ normative predicates. But since I have defined conativism as a theory about states of mind, rather than the meanings of sentences, the Frege-Geach objection applies to conativism only if we express it—as I do in the main text—as an objection that makes reference to the nature of states of mind rather than to the meanings of sentences. (This is not wholly unusual; see, e.g., Unwin, "*Quasi*-Realism, Negation and the Frege-Geach Problem" and "Norms and Negation.") This is plausibly the best way to think about the fundamental source of the Frege-Geach problem: expressivists claim that meanings are inherited from the states of mind they express, and so if we can find states of mind that constitute logically complex normative judgments, it seems likely that we thereby give expressivists the materials they need to explain the meanings of logically complex normative sentences. Interestingly, Mark Kalderon takes the reverse view (*Moral Fictionalism*). He too distinguishes between the psychological theory of noncognitivism, and the semantic theory of expressivism (*Moral Fictionalism*, 52–53 and 95–146). But in contrast to me, he claims that the Frege-Geach problem is generated by expressivism rather than noncognitivism (*Moral Fictionalism*, 52–94). It is for this reason that he claims that his *fictionalist* theory, which commits to noncognitivism but not expressivism, avoids the Frege-Geach problem. But the claim that Kalderon's brand of fictionalism avoids the Frege-Geach problem is mistaken—see Eklund, "The Frege-Geach Problem and Kalderon's Moral Fictionalism," and references therein.

40 Schroeder, *Being For*, 39–55; Unwin, "*Quasi*-Realism, Negation and the Frege-Geach Problem" and "Norms and Negation."

How does this objection relate to the objection I have raised in this paper? The two are distinct, insofar as the examples I have focused on have been logically atomic normative judgments (e.g., the judgment that [Jeff ought to ϕ]), and the Frege-Geach objection appeals to logically complex normative judgments. But though these two problems are distinct, they point to the same conclusion: not all normative judgments are desires.[41] Perhaps there are other respects in which conativism is problematic, but the objection I have raised, and the Frege-Geach objection, both focus on one specific feature of conativism: that it analyzes all normative judgments as desires. Any theory that does not commit to this claim promises to thereby avoid both objections.

7. OTHER THEORIES

I have suggested that there are some grounds for doubt about the motivation argument for conativism, and equal grounds for doubt about conativism itself: other-regarding normative judgments do not seem to have motivational powers, and do not seem to be desires. But this leaves open that we might accept some other, more modest, formulation of the motivation argument and some other, more modest, conclusion.

The problematic judgments that I focused on were other-regarding normative judgments. So we might formulate the motivation argument with reference to self-regarding (*de se*) normative judgments and draw the conclusion that such normative judgments are desires. But we might restrict the motivation argument further. For example, given the Frege-Geach problem, we might doubt that logically complex normative judgments have motivational powers, and doubt that they are desires. So we might formulate the motivation argument with reference to self-regarding logically atomic normative judgments, and draw the conclusion that just those judgments are desires. In fact, I am going to restrict the argument still further, and formulate it with reference to normative judgments with the content [I have reason to ϕ]. This formulation of the motivation argument is more restricted than my arguments warrant. I work with this formulation of the motivation argument primarily for simplicity: the claims that follow are easier to understand if we express them in these simple and positive terms. As it happens, I also think that this is the most plausible way of specifying the class of normative judgments with motivational import, but I shall not rely on or defend that claim here.

That is, in what follows, I address the following argument, which I label the *best motivation argument*:

41 Cf. Finlay, *A Confusion of Tongues*, 130–34.

P1+: All normative judgments with the content [I have reason to ϕ] can motivate.
P2: Only desires can motivate.
So, C+: All normative judgments with the content [I have reason to ϕ] are desires.

Obviously, P1+ is consistent with the claim that other-regarding normative judgments cannot motivate. So too, C+ is consistent with the claim that other-regarding normative judgments are not desires. So the best motivation argument avoids the problems facing the bold motivation argument. But at the same time, the best motivation argument fails to support conativism, which makes a claim about all normative judgments, not just some normative judgments with certain specific contents.

If we think the best motivation argument looks compelling, we might be inclined to reject pure cognitivism, which claims that normative judgments are never desires. Such a view requires that we reject P1+ or P2 above, and those claims seem attractive. But I have also suggested that we should reject conativism, which claims that all normative judgments are desires. So the remaining possibility is that some normative judgments are desires, and some are not. Such a view would be supported by the best motivation argument, and would not be threatened by other-regarding normative judgments.

There may be very many views of this kind.[42] Here I describe just two, to illustrate the kind of view I have in mind. First, we might adopt some kind of desire-as-belief view, such as the following:

Desire-as-belief: To desire to ϕ just is to believe that you have normative reason to ϕ.[43]

Desire-as-belief reduces desires to a particular kind of normative belief. According to desire-as-belief, normative judgments about oneself having reason to do things are desires, but other normative judgments are not desires. Whatever else we might say about desire-as-belief, it should be clear that it would have the attractive features we want. Desire-as-belief is supported by the best motivation argument, since it explains how the relevant normative judgments can motivate

42 See, possibly, Little, "Virtue as Knowledge"; McDowell, "Are Moral Requirements Hypothetical Imperatives?"; McNaughton, *Moral Vision*, 106–17; Pettit, "Humeans, Anti-Humeans, and Motivation"; Price, "Defending Desire-as-Belief"; and Scanlon, *What We Owe to Each Other*, 7–8, 37–49.

43 See Gregory, "Might Desires Be Beliefs about Normative Reasons for Action?"; cf. Humberstone, "Wanting as Believing"; McNaughton, *Moral Vision*, 106–17.

us: because they are desires. It also fits well with the examples I have discussed in this paper, since it says that those normative judgments cannot motivate us and are not desires. Desire-as-belief seems superior to pure cognitivism because it is consistent with the premises of the best motivation argument, and superior to conativism because it is consistent with the fact that other-regarding normative judgments are not desires.

(Given the focus of the paper, I here focus on the advantages of desire-as-belief with respect to other-regarding normative judgments. But given that I above said that this objection to conativism is structurally similar to the Frege-Geach objection, it should be clear that desire-as-belief promises to avoid both problems. Desire-as-belief says that only (some) logically atomic normative judgments are desires, and thereby rightly entails that no logically complex normative judgment can motivate or is a desire. It thereby avoids the Frege-Geach objection. If the best argument for conativism is the motivation argument, and the worst objection the Frege-Geach objection, desire-as-belief is clearly the superior view: it has the former advantage but not the latter cost.)

It is worth very briefly reminding ourselves of one of the main arguments for desire-as-belief.[44] This argument says that we have to adopt desire-as-belief if we want to explain how it is that desires rationalize our behavior. If we think of desires as mere drives, pushing us around, they seem to explain our physical movements in a mechanistic way that does not show our behavior to be rational. In contrast, if we think that our desires represent their objects as normatively favored in some way, it seems that we can explain why it is that desires rationalize, rather than merely cause, our behavior.

This argument not only promises to justify desire-as-belief, but also helps us to respond to a possible objection to the view that arises in the context of this paper. One might worry that desire-as-belief merely *stipulates*, rather than *explains*, the fact that other-regarding normative judgments are not desires. But if desire-as-belief is justified by appeal to the argument above, we can reply to this objection. The argument above suggests that desires are privileged in explaining our behavior not because only they can cause physical movements, but because only they can rationalize, rather than merely cause, our behavior. But if this is so, we can immediately see why other-regarding normative judgments are not desires: because their normative content could not rationalize our own behavior. To this extent, desire-as-belief does promise to explain why other-regarding normative judgments are not desires. No doubt more could be said here, but it is clear that if desire-as-belief is attractive at all, it has the resources to explain why other-regarding normative judgments are not desires.

44 See, e.g., Quinn, "Putting Rationality in Its Place."

It is helpful to approach the overall topic from another direction. Conativism and desire-as-belief might look like very similar theories: conativism reduces normative judgments to desires, and desire-as-belief reduces desires to normative judgments. So the two views might seem similar in that both identify normative judgments and desires with one another. One issue that divides the two views is whether both normative judgments and desires are beliefs, as desire-as-belief says, or whether neither normative judgments nor desires are beliefs, as conativism says. This issue may be difficult to resolve, especially once conativists adopt quasi-realism and claim that normative judgments are beliefs in some minimalist sense.[45] But there is a further and clearer contrast between the two views, and that is the scope of the identity that they posit. Conativism identifies all normative judgments with desires, whereas desire-as-belief identifies only some normative judgments with desires. Here, I suggest that desire-as-belief has the upper hand. For example, one central difference between desire-as-belief and conativism is that desire-as-belief does not treat other-regarding normative judgments as desires. It can thereby better accommodate other-regarding normative judgments while retaining the central attraction of explaining the motivational power of (some) normative judgments.[46]

A second view that says that some but not all normative judgments are desires is what I will call *modest* desire-as-belief. Such a view distinguishes between two different kinds of desire, and analyzes only some as normative beliefs:

> Modest desire-as-belief: Desires are either reflective or brute. To reflectively desire to ϕ is to believe that you have a normative reason to ϕ. To brutely desire to ϕ involves no normative beliefs.[47]

Modest desire-as-belief says that we have two kinds of desire, and that some reduce to beliefs about what we have reason to do and others do not. Such a view also gains support from the best motivation argument. It explains how some normative judgments can motivate us: because they are (reflective) desires. But it also permits that other normative judgments cannot motivate us, as I have argued. In these respects, it is just the same as desire-as-belief. It differs from desire-as-belief in that it permits that not all of our desires involve normative

45 Dreier, "Meta-Ethics and the Problem of Creeping Minimalism."

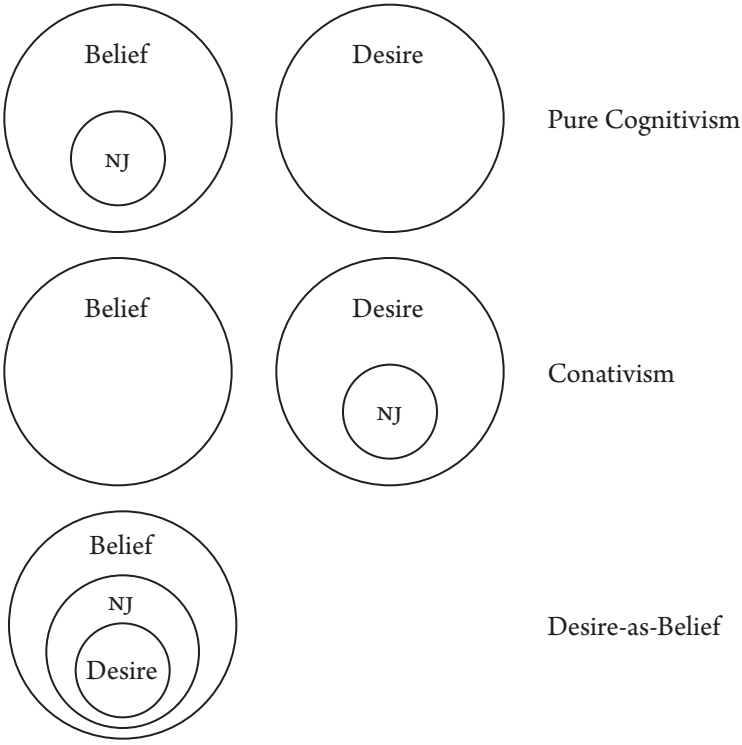
46 Note also that desire-as-belief permits that even some normative judgments with the content [I have reason to ϕ] do not motivate us, and thereby permits weakness of will. It commits only to the claim that such judgments have the power to motivate us. See Gregory, "Might Desires Be Beliefs about Normative Reasons for Action?"

47 Cf. Nagel, *The Possibility of Altruism*; Schueler, *Desire*.

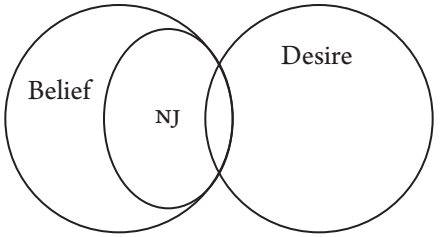
judgments, and this might seem attractive when we think about desires for the bad, appetites, and the desires of animals.[48]

Perhaps other views also share these merits, entailing that some but not all normative judgments are motivational. Like conativism, such views permit that (some) normative judgments have motivational powers that only desires have. But unlike conativism, they also permit that other normative judgments lack those motivational powers. That is, such views, like conativism, have the payoff that they explain the motivational role of normative judgment. But unlike conativism, they avoid committing to the claim that all normative judgments have such motivational powers.

It might be helpful to illustrate the four candidate views as Euler diagrams (using "NJ" as shorthand for "normative judgment"):



48 See, e.g., Stocker, "Desiring the Bad"; and Velleman, "The Guise of the Good." For responses, see, e.g., Raz, "Agency, Reason, and the Good" and "On the Guise of the Good"; and Gregory, "The Guise of Reasons."



Modest
Desire-as-Belief

The first Euler diagram represents pure cognitivism. On that view, there is an exclusive distinction between beliefs and desires, and normative judgments are a subset of our beliefs. The best motivation argument threatens to undermine this view, since it threatens to show that some normative judgments are desires. The second Euler diagram represents conativism. On that view, there is an exclusive distinction between beliefs and desires, and all normative judgments are a subset of our desires.[49] In this paper, I have suggested that this view is also mistaken, since it wrongly treats all normative judgments as desires. That claim is not demonstrated by the best motivation argument, and faces objections from other-regarding normative judgments (as well as from logically complex normative judgments).

The third Euler diagram represents desire-as-belief. According to desire-as-belief, there is no exclusive distinction between beliefs and desires. Rather, our desires are a subset of our beliefs. In particular, they are beliefs with a particular normative content. Since these beliefs are desires, desire-as-belief fits well with the best motivation argument: these beliefs can motivate even though only desires can motivate. But not just any belief with a normative content is a desire: that is how the view is consistent with the examples that limit the scope of the motivation argument, and that cast doubt on conativism. The fourth Euler diagram represents modest desire-as-belief. According to modest desire-as-belief, there is no exclusive distinction between beliefs and desires. Rather, some normative judgments are both. That is how modest desire-as-belief fits well with the best motivation argument: it explains how some normative judgments can motivate us. But again, modest desire-as-belief does not say that all normative judgments are desires, and as such avoids the objections I have been pressing against conativism.

49 Quasi-realist conativists might say that the distinction between beliefs and desires is not exclusive, because normative judgments are both—they are beliefs in some minimalist sense. But this makes no real difference here, since such a view is still committed to the claim that all normative judgments are desires.

8. Conclusion

In this paper I first suggested that the motivation argument fails to support conativism. Only some normative judgments have the capacity to motivate, and this provides inadequate support for the claim that all normative judgments are desires. Second, I suggested that those same problematic normative judgments cast doubt on conativism itself, since such judgments are not plausibly analyzed as desires. Third, I suggested that other views, such as desire-as-belief, are well placed to accommodate the motivational import of normative judgment. They are consistent with the attractive claims that some normative judgments have motivational powers and are desires. But unlike conativism, they allow that not all normative judgments have motivational powers or are desires. I conclude that such theories may hold greater promise than conativism.[50]

University of Southampton
a.m.gregory@soton.ac.uk

References

Blackburn, Simon. *Essays in Quasi-Realism*. Oxford: Oxford University Press, 1993.

———. *Ruling Passions*. Oxford: Oxford University Press, 1998.

———. *Spreading the Word*. Oxford: Oxford University Press, 1984.

Brink, David O. "Moral Motivation." *Ethics* 108, no. 1 (October 1997): 4–32.

Björklund, Fredrik, Gunnar Björnsson, John Eriksson, Ragnar F. Olinder, and Caj Strandberg. "Recent Work on Motivational Internalism." *Analysis* 72, no. 1 (January 2012): 124–37.

Broome, John. "Reasons and Motivation." *Proceedings of the Aristotelian Society, Supplementary Volumes* 71, no. 1 (1997): 131–46.

D'Arms, Justin, and Daniel Jacobson. "Expressivism, Morality, and the Emotions." *Ethics* 104, no. 4 (July 1994) 739–63.

Dancy, Jonathan. *Moral Reasons*. Oxford: Blackwell, 1993.

Davidson, Donald. "Actions, Reasons, and Causes." In *Essays on Actions and Events*, 3–19. Oxford: Oxford University Press, 2001.

50 For useful advice on this paper, I should thank *at least*: colleagues at Southampton (especially Jonathan Way), audiences at Antwerp, Humboldt and Leeds universities, Teemu Toppinen, and two referees for this journal.

Dreier, James. "Meta-Ethics and the Problem of Creeping Minimalism." *Philosophical Perspectives* 18 (2004): 23–44.
Eklund, Matti. "The Frege-Geach Problem and Kalderon's Moral Fictionalism." *Philosophical Quarterly* 59, no. 237 (October 2009): 705–12.
Finlay, Stephen. *A Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press, 2014.
Fletcher, Guy, and Michael Ridge, eds. *Having It Both Ways: Hybrid Theories in Meta-Normative Theory*. Oxford: Oxford University Press, 2014.
Gibbard, Allan. *Meaning and Normativity*. Oxford: Oxford University Press, 2012.
———. *Thinking How to Live*. Cambridge, MA: Harvard University Press, 2003.
———. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press, 1990.
Gregory, Alex. "The Guise of Reasons." *American Philosophical Quarterly* 50, no. 1 (January 2013) 63–72.
———. "Might Desires Be Beliefs about Normative Reasons for Action?" In *The Nature of Desire*, edited by Julien Deonna and Federico Lauria, 201–17. New York: Oxford University Press, 2017.
Hare, R. M. *The Language of Morals*. Oxford: Oxford University Press, 1952.
———. *Moral Thinking*. Oxford: Oxford University Press, 1981.
Humberstone, Lloyd. "Direction of Fit." *Mind* 101, no. 401 (January 1992): 59–83.
———. "Wanting as Believing." *Canadian Journal of Philosophy* 17, no. 1 (March 1987): 49–62.
Kalderon, Mark Eli. *Moral Fictionalism*. Oxford: Oxford University Press, 2005.
Korsgaard, Christine M. "Skepticism about Practical Reason." *Journal of Philosophy* 83, no. 1 (January 1986): 5–25.
Little, Margaret O. "Virtue as Knowledge: Objections from the Philosophy of Mind." *Noûs* 31, no. 1 (March 1997): 59–79.
McDowell, John. "Are Moral Requirements Hypothetical Imperatives?" In *Mind, Value, and Reality*, 77–94. Cambridge, MA: Harvard University Press, 1998.
McNaughton, David. *Moral Vision: An Introduction to Ethics*. Oxford: Blackwell, 1988.
Mele, Alfred R. "Motivation: Essentially Motivation-Constituting Attitudes." *The Philosophical Review* 104, no. 3 (July 1995): 387–423.
Nagel, Thomas. *The Possibility of Altruism*. Princeton: Princeton University Press, 1970.
Parfit, Derek. *On What Matters*. Vol. 2. Oxford: Oxford University Press, 2011.
Pettit, Philip. "Humeans, Anti-Humeans, and Motivation." *Mind* 96, no. 384 (October 1987): 530–33.

Price, Huw. "Defending Desire-as-Belief." *Mind* 98, no. 389 (January 1989): 119–27.

Quinn, Warren. "Putting Rationality in Its Place." In *Virtues and Reasons: Philippa Foot and Moral Theory*, edited by Rosalind Hursthouse, Gavin Lawrence, and Warren Quinn, 181–208. Oxford: Oxford University Press, 1995.

Raz, Joseph. "Agency, Reason, and the Good." In *Engaging Reason: On the Theory of Value and Action*, 22–45. Oxford: Oxford University Press, 1999.

———. "On the Guise of the Good." In *Desire, Practical Reason, and the Good*, edited by Sergio Tenenbaum, 111–37. Oxford: Oxford University Press, 2010.

Ridge, Michael. *Impassioned Belief*. Oxford: Oxford University Press, 2014.

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

Schroeder, Mark. *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press, 2008.

Schueler, George F. *Desire: Its Role in Practical Reason and the Explanation of Action*. Cambridge, MA: MIT Press, 1995.

Smith, Michael. *The Moral Problem*. Malden, MA: Blackwell, 1994.

Snare, Francis. *Morals, Motivation and Convention: Hume's Influential Doctrines*. Cambridge: Cambridge University Press, 1991.

Stalnaker, Robert C. *Inquiry*. Cambridge, MA: MIT Press, 1984.

Stevenson, Charles L. "The Emotive Meaning of Ethical Terms." *Mind* 46, no. 181 (January 1937): 14–31.

Stocker, Michael. "Desiring the Bad: An Essay in Moral Psychology." *Journal of Philosophy* 76, no. 12 (December 1979): 738–53.

Svavarsdóttir, Sigrun. "Moral Cognitivism and Motivation." *Philosophical Review* 108, no. 2 (April 1999): 161–219.

Unwin, Nicholas. "Norms and Negation: A Problem for Gibbard's Logic." *Philosophical Quarterly* 51, no. 202 (January 2001): 60–75.

———. "Quasi-realism, Negation and the Frege-Geach Problem." *Philosophical Quarterly* 49, no. 196 (July 1999): 337–52.

van Roojen, Mark. "Moral Rationalism and Rational Amoralism." *Ethics* 120, no. 3 (April 2010): 495–525.

Velleman, J. David. "The Guise of the Good." *Noûs* 26, no. 1 (March 1992): 3–26.

Williamson, Timothy. *The Philosophy of Philosophy*. Oxford: Oxford University Press, 2007.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.213

CAN THERE BE GOVERNMENT HOUSE REASONS FOR ACTION?

Hille Paakkunainen

THIS ESSAY ARGUES that normative reasons for action are premises in good practical reasoning. In particular, reasons are considerations that nonnormatively well-informed good deliberation takes into account, and if the reasons are decisive, it is part of good deliberation to be moved to act on them in the way that they support. Something like this claim is often quietly observed as a constraint on theorizing about reasons for action, and sometimes explicitly articulated. Mark Schroeder proposes the "Deliberative Constraint" that "one's reasons are the kinds of thing that one ought to pay attention to in deliberating," and suggests that the relative weights of two sets of reasons, *R* and *S*, for *A* to ϕ depend on which of *R* or *S* it is "correct to place more weight on … in deliberation about whether to [ϕ]."[1] Kieran Setiya says that reasons for *A* to ϕ are premises for "sound reasoning to a desire or motivation to ϕ," and sees this as a "harmlessly illuminating" thesis connecting "two things which surely must be connected": reasons for action and practical thinking or deliberation.[2] Jonathan Way says that it is "near platitudinous" that "a reason for you to ϕ must be an appropriate premise for reasoning towards ϕ-ing."[3]

Borrowing Schroeder's term, call the general idea that reasons for action are considerations that good deliberation takes into account and, if the reasons are decisive, issues in action on, the *Deliberative Constraint*. This constraint is not yet a theory of reasons—at least, not in the version I will defend—but a necessary condition compatible with many further views. While the Deliberative Constraint is relatively orthodox, it has not gone unchallenged, and we currently

1 Schroeder, *Slaves of the Passions*, 33, 26, 130, 136–40.

2 Setiya, "What Is a Reason to Act?" 221, 223.

3 Way, "Reasons as Premises of Good Reasoning," 1. For other articulations of similar theses, see, e.g., Raz, *Practical Reasoning*, 5, "The Truth in Particularism," 228, and *From Normativity to Responsibility*, chs. 2, 5; Darwall, *Impartial Reason*, 30–31; Wallace, "Constructivism about Normativity," 19; and Sinclair, "Promotionalism, Motivationalism, and Reasons to Perform Physically Impossible Actions," 649–50, 658.

lack a good sense of why we should subscribe to it, if at all. My aim is to articulate what is right about the orthodoxy, and to explain why recent challenges to it misfire. I argue that if we abandon the Deliberative Constraint, we are left operating with a notion of reasons for action that cannot make sense of reasons' peculiar normativity, and relatedly, cannot play the usual theoretical roles that give questions about the nature and extent of our reasons for action much of their import. We can decide to operate with such a notion of reasons, of course. But we should realize what is thereby sacrificed.

Thoughts in the vicinity of the Deliberative Constraint can seem obviously true, whether explicitly articulated or implicitly adhered to. For example, Horty frames his study of reasons as a study of the logic of reasoning, assuming it as an obvious feature of everyday discourse that reasons are something to focus on in deliberation, provided we are informed of the considerations that are the reasons.[4] Versions of the same assumption inform Williams's famous argument for "internalism" about reasons, Korsgaard's neo-Kantian internalist alternative, as well as many "externalist" views such as those of McDowell or FitzPatrick.[5] These views disagree not on whether reasons are linked to good deliberation, but on whether the agent whose reasons they are must be motivationally capable of undertaking the relevant good deliberation (as for internalists) and, relatedly, what the relevant kind of goodness in deliberation is.[6] Parfit, too, links practical normative reasons to an ideal of rational response to the considerations that are the reasons:

> The rationality of our desires and acts depends on whether, in having these desires and acting in these ways, we are responding well to practical reasons or apparent reasons [i.e., to considerations that, if true, would be reasons] to have these desires and to act in these ways.[7]

This list of authors who assume that reasons are somehow linked to good deliberation or rational response merely scratches the surface. I do not claim that they all would, on reflection, accept the Deliberative Constraint, in the form that

4 Horty, *Reasons as Defaults*.

5 Williams, "Internal and External Reasons"; Korsgaard, "Skepticism about Practical Reason," *The Sources of Normativity*, and *Self-Constitution*; McDowell, "Might There Be External Reasons?"; FitzPatrick, "Robust Ethical Realism, Non-Naturalism, and Normativity."

6 See Paakkunainen, "Internalism and Externalism about Reasons," for a discussion of how the Deliberative Constraint informs much of the internalist/externalist debate. Some views labeled "internalist" deny that reasons are constrained by agents' motivational capacities: see, e.g., Smith, "Internal Reasons," and Markovits, *Moral Reason*, discussed below.

7 Parfit, *On What Matters*, 1:114, 117. Parfit holds similar views about epistemic rationality and reasons for belief.

I will defend. But it is common to hold theses in the vicinity, regardless of one's further views.

However, it is actually highly unobvious why the normative support relations between considerations and the actions (or action types) they support should be linked to good deliberation, or to patterns of rational response to information. Normative reasons for an action are considerations that count in favor the action, at least *pro tanto*. Could there not be normative reasons, *p*, for agent *A* to ϕ in *C*, that are just "out there," counting in favor of A's ϕ-ing—even decisively so—but that *A* should not or perhaps could not take into account, or act *on*, no matter how well-informed and rationally excellent her deliberations? Many moral theorists hold that the facts that ultimately morally justify ϕ-ing are often to be ignored in morally good deliberation. Bernard Williams evocatively called the utilitarian version of this view "government house utilitarianism." The image evoked is that of lack of transparency, from the point of view of ordinary agents' morally good deliberations, to the facts about utility-maximization that ultimately morally justify individual action and social policy.[8] If morally good deliberation can ignore ultimate moral justifications, why could rationally excellent, well-informed deliberation not ignore normative reasons, and focus on some other considerations entirely? If there can be "government house" moral justifications, why not also "government house" reasons (GH reasons)?

By GH reasons, I mean reasons that violate the Deliberative Constraint, in whatever its best formulation. While the analogy with GH moral theory is loose, it serves to point out that, if some version of the Deliberative Constraint holds, it is not obvious why.[9] Putative examples of GH reasons cast further doubt on the Deliberative Constraint. Consider this well-known case from Mark Schroeder:

> *Surprise Party*: Nate … hates all parties except for successful surprise parties thrown in his honor [which he loves]. Given Nate's situation, the fact that there is a surprise party waiting for him now at home is a reason for him to go home. But it isn't a reason that Nate could know about or act on. Still, someone Nate trusts might [truly] tell him that there is a reason for him to go home now.[10]

The thought is that, although the fact, *p*, that a surprise party awaits is a reason for Nate to go home, Nate cannot believe that *p* without destroying *p*'s status as a

8 Williams, "Ethics and the Limits of Philosophy," 108–10.

9 The statement of GH moral theory could also be refined, e.g., to account for how ultimate or "fundamental" moral justifications relate to "derivative" ones (Star, *Knowing Better*). Section 5 below addresses parallel complications regarding reasons for action.

10 Schroeder, *Slaves of the Passions*, 33.

reason.[11] (The surprise would be ruined, and *ceteris paribus*, Nate has no reason to go to ruined surprise parties, since he hates them.) Since taking *p* into account in deliberation requires believing that *p*, *p* is a reason for Nate only if Nate *does not* take *p* into account in deliberation, and so only if Nate does not take *p* into account in *good* deliberation.[12]

Julia Markovits suggests many further examples:

> *James Bond:* Let's say I become convinced I am James Bond. The fact that I am suffering from such a delusion may give me an excellent reason to see a psychiatrist for treatment. But it cannot motivate me to see the psychiatrist. For if this fact could motivate me to seek help, I would no longer be convinced I was James Bond.[13]
>
> *Soldier in a Just War*: In a war fought on humanitarian grounds, soldiers may have reason to desensitize themselves to the common humanity of the inhabitants of an enemy state so that they can more effectively fight a war whose very justification is provided by that common humanity. If they have reason to fight in the war, and fight effectively, then they ought not to be motivated to fight by that reason.[14]
>
> *Emergency Landing*: Captain Sullenberger successfully emergency-landed an Airbus A320, which had lost all thrust in both engines …, in the icy waters of the Hudson River, with no loss of life. [Asked] whether he had been thinking about the passengers as his plane was descending rapidly …, Captain Sullenberger replied, "Not specifically.… I mean, I knew I had to solve this problem. I knew I had to find a way out of this

11 I will assume that reasons are facts or true propositions, *p* (I will not worry about which). Following convention, I often call both "considerations." For ease of writing, I sometimes say that *A* "has" a reason to ϕ, meaning that there is a reason for *A* to ϕ. I assume that, whenever there is a reason, there is some (possibly conjunctive) consideration, *p*, that is that reason.

12 Ironically, Schroeder introduces the case right after articulating his "Deliberative Constraint" (*Slaves of the Passions*, 26). He does so to argue that existential normative facts such as "there is a reason for me to go home now" might themselves be reasons, since Nate would be deliberating well if he took this fact into account (32–33). But cf. Schroeder, *Slaves of the Passions*, 167.

13 Markovits, Moral Reason, 41. The case is from Johnson's "Internal Reasons," 575. See Smith, "Reasons with Rationalism after All," 523; Millgram, "Williams' Argument Against External Reasons"; and Sobel, "Explanation, Internalism, and Reasons for Action," for similar cases.

14 Markovits, *Moral Reason*, 47

> box I found myself in.... My focus at that point was so intensely on the landing ... I thought of nothing else."[15]

In *James Bond*, as in *Surprise Party*, the agent *cannot* take the putative reason into account without destroying it (perhaps by destroying the fact, *p*, that was supposed to be the reason). In *Soldier* and *Emergency Landing*, the agent *should not* take the putative reasons into account, or be moved by them. Thus it seems that it cannot be part of good deliberation to take them into account and be moved by them, even if the reasons are decisive.[16] Lest we brush such cases aside as rare exceptions, Markovits suggests further everyday examples:

> A specialist ... may be able to cure more patients if she's in it for the social prestige than if she's in it chiefly to save lives.... A surgeon may operate more successfully ... if she is not thinking of the life that is at stake. [We are often fortunate to be] driven by ulterior motives, habit, instinct, or "auto-pilot" rule-following to make decisions or react to threats which we would have likely reacted to less well if we had been responding motivationally to our reasons.... If a child runs into the street right in front of my car, I hit the brakes automatically—I am not motivated by a concern for the well-being of the child. In a surprising number of cases, there is much to be said for *not* being motivated by our reasons.[17]

Such cases look to challenge the Deliberative Constraint. If the Constraint is nonetheless defensible, we should explain why these examples do not defeat it, in its best formulation.

Notice that, even if we abandon the Constraint, there are other ways of linking agents' reasons to their subjective perspectives, or to some hypothetical rational response, that the above examples do not challenge. For example, Markovits's own view (which she calls "internalism") is that a reason for *A* to ϕ is a consideration that counts in favor of *A*'s ϕ-ing "in virtue of the relation it shows ϕ-ing to stand in to [*A*'s] existing ends"—for example, a causal or constitutive means-ends relation, or that of being valuable because of the value of *A*'s ends.[18] Differently, on Michael Smith's "Advice Model," there is a reason for *A* to ϕ if and only if *A*'s perfectly rational counterpart, *A*+, who is thinking about what the actual, less-than-perfectly rational *A* should do, would desire that *A*

15 Markovits, *Moral Reason*, 48.

16 Although, as I explain in section 2 below, this inference is suspect.

17 Markovits, *Moral Reason*, 48–49.

18 Markovits, *Moral Reason*, 52. Cf. Schroeder, *Slaves of the Passions*.

ϕs.[19] Crucially, *A+* might desire that A ϕs, even if there is no good deliberative route *via* which *A* might come to do ϕ on the basis of her normative reasons for ϕ-ing. We should explain why it is that, even if we link agents' reasons to their subjective perspectives or to hypothetical rational responses in these other ways, we still miss out on something important if we abandon the Deliberative Constraint.

I proceed as follows. Section 1 formulates the version of the Deliberative Constraint I defend, and explains why it is relatively modest and available to many theorists. Section 2 defends the Constraint against the counterexamples above. Sections 3–5 develop a positive argument for the Constraint, while responding to objections and introducing some qualifications. The core of the positive argument is that, despite its modesty, the Constraint captures something deeply important about reasons for action: an aspect of their relationship to agents without which we cannot make sense of reasons' peculiar normativity. We can choose to operate with a notion of reasons that abandons the Constraint, but only at the cost of failing to make sense of how reasons can fulfill certain familiar and important theoretical roles associated with their normativity. Leaving these roles behind and hewing to others, we could operate with a different reason-concept, provided we are clear that that is what we are doing. Still, the Constraint is a condition of making sense of reasons' peculiar normativity. Section 6 concludes.

1. FORMULATING THE DELIBERATIVE CONSTRAINT

To formulate the Deliberative Constraint, let us start with Setiya's version, revising as problems arise:

> *Setiya's Reasons* (SR): The fact that *p* is a reason for *A* to ϕ [in circumstance *c*] just in case *A* has a collection of psychological states, *C*, such that the disposition to be moved to ϕ by *C*-and-the-belief-that-*p* [in *c*] is a good disposition of practical thought, and *C* contains no false beliefs.[20]

Some initial clarifications: first, "practical thought" is meant inclusively. ϕ-ing because *p*, where *p* is a consideration on which one acts, is a limiting case of practical thought or deliberation whose premise is *p*. More complex phenomena, such as forming beliefs about which considerations, $p_1 \ldots p_n$ are reasons for

19 Smith, *The Moral Problem*, 156–61, "Internal Reasons," and "Reasons with Rationalism After All." Smith emphasizes that his Advice Model is unthreatened by examples such as those that challenge the Deliberative Constraint ("Reasons with Rationalism after All," 523–24; cf. Sobel, "Explanation, Internalism, and Reasons for Action").

20 Setiya, *Reasons without Rationalism*, 12, and "What Is a Reason to Act?" 222.

what, and acting on the basis of such beliefs, also count.[21] Second, reasons are facts or true propositions *p*, and these are often ordinary nonnormative facts (e.g., [Bert is in pain]). It is the further fact that [the fact that *p* has the property of being a reason] that is a normative fact.[22] SR states the putative conditions under which such normative facts obtain. Third, the final clause, that *C* contains no false beliefs, is designed to rule out cases in which otherwise impeccable reasoning that depends on false beliefs leads one to do something one intuitively has no reason to do. I would like a gin and tonic; I falsely believe that this glass contains gin when it contains petrol; and I am led by cogent means-ends reasoning to mix the contents with tonic and drink it. I have no real reason to drink the mixture.[23] Normative reasons correspond only to deliberation whose course does not depend on false beliefs, and in this sense to "sound" rather than merely "valid" deliberation.[24]

I accept these clarifications—though I call "deliberation" what Setiya calls "practical thought"—and I defend merely a conditional claim, not a biconditional, for reasons I will explain. For now, note two problems with SR's focus on motivational dispositions.

First, sometimes we have reasons for various conflicting act options. Seeing this, Setiya suggests that, because SR is concerned with *pro tanto* reasons that might be overweighed, we should not connect A's reasons to A's ϕ-ing or intending to ϕ as a result of good deliberation, but only to A's having a "motivation or desire" to ϕ, as SR does.[25] This is fairly plausible. When reasons to ϕ are overweighed by reasons to ψ, sound deliberation surely would not lead A to (intend to) ϕ. But likewise, why think that good deliberation in the face of reasons for conflicting options must lead one to have a "desire or motivation"

21 Setiya, "What Is a Reason to Act?" 221. Cf. Way, "Reasons as Premises of Good Reasoning," 2–3, on the broadness of "reasoning."

22 Cf. Dancy, "Nonnaturalism," 137; Parfit, *On What Matters*, vol. 2, 330–31.

23 Williams, "Internal and External Reasons," 102–3.

24 More carefully, the type of dependence on false belief to rule out is (i) dependence on false belief-contents as premises in deliberation, and (ii) distorting effects of false background assumptions on how one deliberates. Deliberation *can* be sound while depending on the presence of false beliefs in certain other ways. A may have reasons to rid herself of false beliefs, and can deliberate soundly about how to do so, aptly relying, in the process, on the true belief that she has false beliefs. Sound deliberation can also sometimes depend on *ignorance* of fact—as when one has reasons to inquire into a topic, or differently, when deciding which horse to bet on (cf. Setiya, "What Is a Reason to Act?" 224). In such cases, one's reasons are facts other than those facts *about* which one is ignorant or mistaken. (Among these other, reason-giving facts might be the fact *that* one is ignorant about *X*, or harbors erroneous beliefs about *X*.)

25 Setiya, "What Is a Reason to Act?" 222n5.

for each option (presumably via the activation of corresponding motivational dispositions)? Reasons may be considerations that good deliberation somehow *registers*. But better to hold that good deliberation can register reasons by weighing them, at least in rough comparative terms, where this need not involve activating or engendering conflicting motivations.[26] Plausibly, it is part of good deliberation to be motivated by the weightiest reasons. But motivation need only accompany the "winning" side.

Second, thinking of registering reasons in terms of weighing, instead of in motivational terms, also helps to distinguish reasons from enabling or disabling conditions. The fact, *p*, that I can keep my promise is not a reason to keep it, but an enabling condition on other facts' being reasons to keep it.[27] Still, it may be good to note in deliberation that I can keep the promise (suppose I previously thought I cannot), and this may make the difference between having and lacking the (good) motivation to keep it. Setiya admits that SR counts *each* fact, *p*, such that the disposition to be moved to ϕ by *C*-and-the-belief-that-*p* is a good disposition of practical thought, as a reason to ϕ; and so that it cannot distinguish reasons from enablers.[28] In contrast, good deliberation plausibly would not assign to enablers (or disablers) weights and valences they lack.[29]

One might object that it is overly intellectualized to construe good deliberation as involving thought about the comparative weights of reasons, even implicitly. Further, sometimes it takes too long to register *all* of our reasons *pro* and *con* all the different act options. Surely we often deliberate as well as needed while only registering the most important reasons in the situation, as when action is needed soon.[30] The right response here, I think, is to accept these claims, but to accommodate them in our understanding of the Deliberative Constraint.

In response to the charge of overintellectualizing: registering the comparative weights of reasons need not involve beliefs about reasons and their weights, considered as such. It might instead be a matter of how one updates one's (conditional) preferences concerning one's act options, upon considering the reason-giving facts, *p*, along with other relevant facts such as enablers or disablers.

26 Doubts about "intrinsic masking" supply one reason not to ascribe conflicting motivational dispositions (Handfield and Bird, "Dispositions, Rules, and Finks"; cf. Ashwell, "Superficial Dispositionalism," and Clarke, "Opposing Powers"). But it is also unintuitive that good deliberation in the face of reasons for conflicting options must involve conflicting motives; cf. Schroeder, *Slaves of the Passions*, 166.

27 Dancy, *Ethics without Principles*, 38–41.

28 Setiya, "What Is a Reason to Act?" 226.

29 In allowing disablers, the weighing conception may also have an advantage over Way's view ("Reasons as Premises of Good Reasoning," 268n25).

30 Sobel, review of *Slaves of the Passions*.

For instance, registering a disabler—say, that in the circumstances, I cannot keep a promise I made—might leave in place a strong conditional preference for keeping my promise, should it turn out that I can keep it after all.[31] Registering decisive reasons against keeping the promise (perhaps keeping it would have disastrous consequences) would remove this conditional preference in a good deliberator. The details are work for theorists of deliberation. The general idea of weighing reasons in deliberation is intuitive enough.

Relatedly, regarding the question of time: we register and act on reasons all the time, often quickly and fairly automatically. Seeing a child running onto the road, we hit the brakes, and the fact that the child ran onto the road was a reason—and a decisive one—to do so.[32] Such quick responses can be cases of acting on reasons, and so of "deliberation" in the broad sense identified above. (Compare the fastness and automaticity of most inferences involved in everyday conversations, or in simple arithmetic.) Indeed, it is part of the point of training in a complex activity that one's skilled, trained responses become fairly fast and automatic, while remaining intelligent, rational responses to key reason-giving features of situations. We need not in any case claim that *every* case of good deliberation involves responding to *all* of one's reasons. Rather, whenever p is a reason for or against A's ϕ-ing in circumstance C, there is a possible course of good deliberation in C that does involve A's assigning to p the (comparative) strength, s, and valence, v, that p has in C. This leaves room for instances of good deliberation that are more truncated.

This last formulation is roughly correct, but let us explicitly forestall two confusions. First, it often is not good deliberation to respond to only *one* or *few* of the reasons in a situation, especially if the reasons responded to are insignificant while the reasons ignored are weighty. Suppose that if I marry Fred I will be a dollar richer than if I marry Frida, and suppose that this fact is a reason, albeit a very insignificant one, to marry Fred rather than Frida. I am aware of other relevant facts that are weighty reasons for or against marrying either, but I ignore those in my decision-making, making my decision solely based on the extra dollar. This would not plausibly be good deliberation. There are limits to how truncated good deliberation can be. Good deliberation in a situation cannot focus exclusively on insignificant reasons while ignoring significant ones. This does

31 This need not involve having an unconditional motivation or desire toward keeping the promise.

32 What about the fact that the child's well-being is at stake—a fact we might have no time to register? (Markovits, *Moral Reason*, 48–49) This raises the question of which of many nearby facts to count as "the" reason or reasons; as well as which reasons are "fundamental" and which ones are "derivative." Section 5 discusses these issues.

not conflict with the formulation above, but it is worth making explicit. Well-informed good deliberation should at least register *all of the important* reasons in the situation, lest it lead one astray from what the balance of reasons supports.

Second, and related, we should explicitly rule out the following possibility: that although there is, for each (important) reason *p*, *some* possible course of deliberating well in *C* that assigns to *p* the (comparative) strength and valence that *p* has in *C*, there is no *single* possible course of deliberating well that takes into account *all of the important* reasons, $p_1 \ldots p_n$ in *C*. Again, well-informed good deliberation should intuitively respond to all of the important reasons in the situation, lest it lead the agent astray from what the balance of her reasons supports.

Incorporating these points, and henceforth restricting our attention to important (in the sense of relatively weighty in the context) reasons, here is a revised version of the Deliberative Constraint:

> *Reasons Revised* (RR): The fact that *p* is a reason of strength *s* for *A* to ϕ in *C* only if there is a course of deliberating well *tout court* such that *A* could undertake it in *C*, therein taking *p* into account (along with other relevant facts such as further reasons, enablers, and disablers) and weighting *p*'s strength as *s*—where her so weighting *p* depends on no false beliefs. Deliberating well concludes either in acting on the weightiest reasons, in the way that they support, or in forming an intention to so act, or, if no set of reasons is weightiest, in true belief about which actions are the permitted or best options.[33]

With some qualifications introduced in section 5, RR is the version of the Deliberative Constraint I defend. If RR holds, then at least our important reasons cannot be GH reasons. And it is unclear why we should posit the existence of GH reasons at all, if they are doomed to unimportance as normative phenomena.[34]

Four final clarifications before arguing for RR. First, RR is proposed as a conceptual truth about reasons. But if RR is true and our concept of a reason is not erroneous, our metaphysical views about the nature of reasons and of good deliberation should respect the link that RR records.[35]

33 Three quick notes: (1) Strictly, RR also concerns reasons *against*, which are to be taken into account and weighted. For simplicity, the formulation omits reference to these (and to valences), speaking only of reasons *for*. (2) I clarify below the sense in which RR requires that *A* "could" undertake the relevant course of good deliberation. (3) When is no set of reasons weightiest? E.g., when one's reasons are incommensurable, equally good, or "on a par" (Chang, "Introduction").

34 Recall that reasons can be conjunctive facts. If many insignificant reasons together constitute an important reason, then RR applies to it.

35 Sections 3 and 6 discuss the possibility of operating with different concepts or conceptions

However, second, many different views could respect the proposed conceptual truth. RR links reasons not to deliberation that is good in some qualified sense (instrumentally, morally, prudentially), but to deliberation that is good *tout court*, without qualification. RR does not say what such deliberation is like, aside from involving weighting reasons in accord with their actual strengths, and acting on the weightiest reasons. Deliberation that is good *tout court* may also turn out to be, say, instrumentally or morally good, but RR does not say so. Further, RR is silent on whether reasons' status and weight as reasons is grounded in their status as premises in good deliberation, or vice versa—or neither.

Third, on the idea of deliberating well *tout court*: I said this involves weighting reasons in accord with their actual strengths. The idea of deliberating well *tout court* therefore makes reference to the concept of a normative reason. Does this make RR either trivial and unhelpful, or objectionably circular? No: circularity is a problem for views with reductive ambitions. RR is offered as a conceptual truth, but *not* as a reductive definition of the concept of reasons for action in terms of the concept of deliberating well *tout court.* The two concepts can be interconnected yet irreducible to one another.[36] Nor is RR trivial. As noted (in the introduction), we can reasonably doubt whether normative support relations between reasons and actions are necessarily linked to rational responses to the reason-giving facts. And we can raise putative counterexamples. Even if the idea of deliberating well *tout court* is the idea of rationally responding to, and acting on, normative reasons, RR may be false: there may be reasons for which no course of deliberating well *tout court*, in the relevant sense, is available.[37]

Finally, a clarification about the sense in which RR requires that *A* "could" undertake a course of good deliberation that takes accurate account of her reasons. RR does not assume that *A* has the present actual motivational capacity to do so. The closest worlds in which *A* undergoes such a course of deliberation may be ones in which *A* acquires the motivational capacity to do so—perhaps by acquiring dispositions she actually lacks. This is not to say that *A* might have *no* capacity for deliberation, and yet have reasons for action. Plausibly, only minimally rational agents—agents capable of acting on the basis of considerations, and so of "deliberating" in the broad sense introduced—can have normative reasons for action. But RR leaves it open that *A* can have reasons while lacking the capacity

of reasons.

36 Familiarly, however, conceptual irreducibility leaves room for ontological reducibility (cf. Schroeder, "Realism and Reduction," and *Slaves of the Passions*, ch. 4).

37 If there were generally no such thing as responding rationally to, and acting on, the reason-giving facts, we might conclude that there is no philosophically interesting concept of deliberating well *tout court* to be had. But this conclusion would be wildly premature here.

to go through the specific deliberative routes to which her reasons correspond. *A*'s deliberating well in the relevant way need not be motivationally possible for *A*, but only metaphysically possible, holding fixed facts about the nature of deliberation, about *A*'s being a minimally rational agent, and about what deliberating well in circumstance *C* is like.[38]

RR thus does not presuppose "internalism" about reasons, understood as the view that necessarily, if p is a reason for A to ϕ in C, then A is capable of being motivated to ϕ by the belief that p (together with whatever else—e.g., preexisting desires—is needed for such motivation).[39] Still, RR *also* leaves it open that such internalism *is* a further necessary condition on reasons. I formulate RR as a conditional claim, not a biconditional, to preserve its modesty in this regard. While I only argue for RR—thus securing the claim that reasons are premises in good deliberation—section 4 briefly considers whether my argument might extend to support internalism.

RR is a version of the Deliberative Constraint that is relatively modest and compatible with many different views, yet unobvious and nontrivial.[40] Why believe it—especially given the seeming examples of GH reasons in the introduction? Section 2 defends RR against these examples. In each case, the examples are naturally interpretable as consistent with RR, and the motivations for an anti-RR interpretation are at least as theoretically contentious as RR itself. Sections 3–5 then mount a positive argument for RR, giving a principled reason for interpreting the seeming counterexamples in an RR-friendly way.

2. THE COUNTEREXAMPLES

Recall *Surprise Party*. The fact that a surprise party awaits at home is supposed to be a reason for Nate to go home, but it is a fact that Nate cannot believe without destroying its status as a reason. So, the thought goes, it is a reason that Nate cannot take into account in good deliberation. *Mutatis mutandis* for *James Bond*. What to say about these putative counterexamples to RR?[41]

It is in fact unobvious precisely what normative or evaluative phenomena

38 It is of course a good question what contingencies of *A*'s psychology may be part of "*C*." Cf. note 46 below.

39 Paakkunainen, "Internalism and Externalism about Reasons," discusses this and other versions of internalism. See also the essays in Setiya and Paakkunainen, *Internal Reasons*; and Lord and Plunkett, "Reasons Internalism."

40 RR is also neutral on "particularism" about reasons à la Dancy, *Ethics without Principles*.

41 Traditional "conditional fallacy" examples (Johnson, "Internal Reasons and the Conditional Fallacy") do not threaten RR. RR allows that one's being a generally bad deliberator can itself be a reason, e.g., to improve one's deliberative skills. Satisfying the right-hand side of

our intuitions are tracking in such cases.[42] In *Surprise Party*, perhaps the fact that a surprise party awaits is a reason for Nate to be glad if he ends up getting surprised: *a reason for an affective response, should a pleasant outcome occur*. Or it might be an *explanatory reason*: a fact that explains why, say, Nate's going home would be a favorable outcome from the perspective of Nate's preference-satisfaction; or why Nate's going home would be good, or good for him. Or it might be *a normative reason for Nate's friends* to urge Nate to go. Further, the fact that Nate would be glad if he went, or that it would be good for him, might itself be a normative reason for Nate to go, a reason that Nate *can* take into account and act on: RR does not conflict with these claims. Various possible distinct evaluative and normative phenomena are in the vicinity and, further, *various candidate facts that might be normative reasons for Nate to go*, besides the specific fact that a surprise party awaits. It is unclear that our intuitions in the case track the existence of this specific normative reason for Nate to go home, rather than (or as well as) some of these other normative or evaluative phenomena. Likewise for *James Bond*.

Accordingly, I doubt that such examples are decisive on their own, even if multiplied. Whether our intuitions in these cases track the existence of GH reasons for action instead of (or as well as) other evaluative or normative phenomena is partly to be decided on theoretical grounds: on grounds of whether, e.g., an agent's reasons for action are a function of her preference-satisfaction and, if so, what kind of function. Such theoretical commitments are at least as contentious as RR.

Of course, we will want a positive argument for RR, and so for interpreting our intuitions in these cases compatibly with it. Sections 3–5 give such an argument, returning in due course to reconsider *Surprise Party*–style cases. For now, what about cases such as *Soldier* and *Emergency Landing*? In *Soldier*, the soldier's putative reason to fight is the fact that the enemy inhabitants share a common humanity, but she should not think about this fact or be moved by it in the midst of fighting, lest she lose her nerve and fight (perhaps dangerously) ineffectively. In *Emergency Landing*, Sullenberger supposedly should not think about the fact that lives are at stake as he is landing the plane, although this fact is a decisive reason to attempt emergency landing. Do these cases challenge RR?

I doubt it. In *Soldier*, the common humanity of the enemy state's inhabitants

---

RR does not guarantee that one is already a generally good deliberator, so does not eliminate the reason. Cf. Setiya, "Introduction," 13–15.

42 Cf. Sinclair, "Promotionalism, Motivationalism, and Reasons to Perform Physically Impossible Actions," 657–58, and "On the Connection between Normative Reasons and the Possibility of Acting for those Reasons," 1214–16; as well as Setiya, "Reply to Bratman and Smith," 538, and "What Is a Reason to Act?" 267n14.

may be only a reason *to enter or join the war*, not a reason *to perform this combat maneuver as opposed to that*. Likewise, in *Emergency Landing*, the lives at stake may only provide a decisive reason to *attempt emergency landing*, not *to perform this flight maneuver as opposed to that*. It *would* intuitively be good deliberation for the soldier to enter the war, or for Sullenberger to attempt an emergency landing, based on the reasons for doing so. But plausibly, when already in the midst of fighting a just war, or of attempting an emergency landing, one should not anymore focus on these initial reasons, but rather on facts relevant to which specific maneuvers to perform. This is not a problem for RR: RR links reasons *in circumstance C* to good deliberation in *C*, and the circumstance of being about to enter a just war differs from that of being in the midst of fighting one. Further, where the common humanity of the enemy inhabitants *is* relevant to which combat maneuvers to perform—as it surely can be, lest one risk brutalizing those inhabitants— intuitively it *would* be good deliberation for the soldier to consider it in her decision-making.[43] In general, it *would* intuitively be good deliberation to choose specific combat or flight maneuvers based on the reasons for those maneuvers, rather than on the basis of some other considerations (say, about today's crossword puzzle) that do nothing to justify those maneuvers. Indeed, it is unclear why we should expect agents to choose the right maneuvers *without* taking into account the reasons for those maneuvers, or by taking into account some other facts entirely.[44]

Finally, recall that responding to reasons is often fairly quick and automatic, especially in seasoned performers of an activity. Taking account of relevant reasons even in the midst of complex activity need not excessively distract, nor be the object of unnerved attention. Some *may* get unnerved if they think about the relevant reasons. But we need not say that it is part of good deliberation for this to happen. It is only if one's decisions issue from bad deliberation or nonrational influence, such as being unnerved or losing focus, rather than from responding rationally to reasons in the ways that they support, that one's belief in the reason-giving facts will likely cause one to perform maneuvers one should not perform. The fact that one could get unnerved is no reason to doubt RR.[45]

43 Markovits admits to such a risk in *Moral Reason*, 47n24.

44 The soldier's normative reasons for performing a maneuver may include the fact that she was commanded to do so: the soldier's reasons need not coincide with her superiors' reasons for commanding these maneuvers. I leave aside these (first-order normative) complications here.

45 Compare: it is no part of Williams's ("Internal and External Reasons") "internalism" to deny that agents might fail, through getting unnerved, to go through the good deliberative routes to which their reasons correspond. The above points also account for the further everyday examples Markovits proposes in *Moral Reason*, 48–49.

The foregoing raises two general points worth spelling out explicitly. The first concerns so-called "nonbasic" actions, their nature as processes that unfold over time, and concurrent deliberation as a mechanism of rational control over the process. (Nonbasic actions are actions done *by* doing something else: e.g., landing a plane, by performing certain maneuvers.) While section 1 did not discuss this possibility, RR allows that, even as *A*'s deliberation concludes in her initiating some action, A may keep deliberating further about how, precisely, to carry out the action. We just need to understand RR as applying at different levels—to the *whole* action (e.g., landing the plane) and the decision to perform it, and again to parts of the whole, and the decisions to perform those parts, as the situation evolves. When a complex, skilled performance is in progress, we quickly and fairly automatically choose subactions as ways of performing the whole, rapidly adjusting our performance to the circumstance in an exercise of continuous rational control over our behavior.

The second point concerns rational versus nonrational responses to reason-giving facts, and whether it is sometimes better if agents do not deliberate in light of those facts. As noted, the possibility of getting unnerved if one considers the reason-giving facts does not tell against RR. But suppose one is *very likely* to get unnerved, and very unlikely to respond to one's reasons rationally. Perhaps it is better here not to deliberate at all, or to only consider some other facts. However, we need not deny that it may be in *some sense better* if some agents do not deliberate in some circumstances, or do not take into account reason-giving facts. Even if it would not be *good to deliberate*, or good to deliberate on the basis of the reason-giving facts, because one is unlikely to respond rationally to them, it can still be a case of *deliberating well* to respond to the reason-giving facts in the way that they support.[46]

In sum, *Soldier* and *Emergency Landing* do not defeat RR either. They are plausibly interpreted in line with RR. In each case, it is intuitively a case of good delib-

46 Perhaps you may even have decisive reasons not to deliberate (suppose an evil demon says she will destroy the world if you do). Still, if you also have decisive reasons for something else, ϕ, besides refraining from deliberating, then there is a possible course of good deliberation—that you have decisive reasons not to undertake!—that takes your reasons to ϕ into account and leads you to ϕ. Thanks to André Gallois for pressing me on this. Note that sometimes *A*'s psychological frailties affect not (just) *A*'s likelihood of undergoing the good deliberative routes to which *A*'s reasons correspond, but *A*'s reasons themselves—perhaps by affecting the circumstance in which certain facts are reasons for *A*. (Cf. the sore loser example discussed in Smith, "Internal Reasons.") I doubt there is a good general formula for when psychic frailties affect *A*'s reasons, by affecting the circumstance, and when they affect only *A*'s likelihood of abiding by her reasons.

eration to respond to the reason-giving facts by acting on them, in the way they support. Still, what principled argument is there for RR?

3. THE DISTINCTIVE POINT OF APPEALS TO REASONS FOR ACTION

In responding to *Surprise Party*–style cases, I observed that they may involve various different normative or evaluative phenomena, and it is unclear why we should interpret our intuitions as indicating the presence of the specific putative GH reason, rather than in some other, RR-friendly way. Of course, opponents of RR can try stipulating that the GH reasons *are* present. But stipulations can fail. To make progress, we should consider what is distinctive about normative reasons for action, as compared to other normative or evaluative phenomena, such that we should interpret our intuitions in these cases one way rather than another. What is the distinctive point of appeals to reasons for action?

"Normative reason for action" is to some extent a term of art. We can use it as we wish, if we are clear about how we are using it. But there are key theoretical roles that the concept of reasons for action is often asked to play—roles associated with the peculiar normative importance of reasons. My argument for RR will be that if we deny it—or whatever exactly is the best version of the Deliberative Constraint—then we are left operating with a concept of reasons for action that cannot play these roles. We may choose to operate with such a concept. But we should be clear about what is thereby sacrificed. Further, the concept of reasons that *can* play the roles in question is clearly worth theorizing about, and in terms of. It is this concept of reasons of which, I claim, RR holds.

Which theoretical roles are in question? The first is reasons' role in prosecuting questions about what is often called "normative authority" or "robust normativity." All norms, including rules of games or clubs, have "norm-relative" normativity: one can behave better or worse by their lights.[47] But some norms matter more than others. Some have genuine authority over (all or some of) us, others do not. And questions about normative authority seem to be, at least partly, questions about reasons for action. For instance, morality's authority on *A* is usually taken to depend on whether *A* has any (reasonably weighty) reason to do what morality requires.[48] If no one had any reason to do what morality requires,

47 "Norm-relative normativity" is Finlay's term ("Recent Work on Normativity," 332). McPherson ("Against Quietist Normative Realism," 232–33) calls it "formal normativity," in contrast to "robust normativity" or "authority." Copp ("Moral Naturalism and Three Grades of Normativity," 255–58) calls it "generic normativity."

48 It is also usually supposed that morality is *categorically* authoritative only if everyone regardless of their contingent desires has reasons to be moral. For versions of the view that

nor to avoid what is morally prohibited, then it seems that moral requirements and prohibitions would not really matter, would not have genuine authority over us. On the other hand, insofar as *A* does have (reasonably weighty) reasons to do what is morally required, moral requirements have genuine authority over *A*.[49] *Mutatis mutandis* for other norms, such as norms of etiquette or laws. In sum, the following is part of what philosophers usually mean by "normative authority" (as a property of norms or standards, not of individuals or institutions):

> *Reasons Centrism*: Any given norm *N* is authoritative on agent *A* if and only if *A* has some (reasonably weighty) normative reason to do what *N* says that it would be, in some sense, proper or called for (e.g., morally required, required by etiquette or law) for *A* to do.

Nor is this mere philosophers' fiction. It is a regimentation rooted in common human concerns expressed in questions such as "Why be moral?" or "Why obey the rules?"

Reasons Centrism is fairly modest. It does not entail *Reasons Basicness*, the view that the property of being a reason is not analyzable in terms of other normative or evaluative properties, and is what all other normative or evaluative properties are analyzable in terms of.[50] Further, both *Reasons Centrism* and *Reasons Basicness* are silent on whether the property of being a reason is reducible to something nonnormative or nonevaluative, and on whether it is a natural or a nonnatural property.[51]

Still, *Reasons Centrism* is a central organizing assumption of much contem-

---

everyone has such reasons, see, e.g., Shafer-Landau, "A Defence of Categorical Reasons"; Korsgaard, *The Sources of Normativity* and *Self-Constitution*; and Foot, *Natural Goodness*.

49 This authority might be merely hypothetical, if the reasons depend on A's contingent desires.

50 See, e.g., Schroeder, *Slaves of the Passions*, ch. 4; and Skorupski, *The Domain of Reasons*, for *Reasons Basicness*.

51 See, e.g., Scanlon, *Being Realistic about Reasons*; Parfit, *On What Matters*, vols. 1–2; and Enoch, *Taking Morality Seriously*, for nonreductive views, and Schroeder, *Slaves of the Passions*, for a reductive view. The questions of reducibility and naturalism are distinct, if there is room for nonreductive naturalism à la Sturgeon, "Ethical Naturalism." Note also that *Reasons Centrism* does not entail that morality's authority depends on *non*moral reasons to be moral. The relevant reasons might be moral reasons, whatever this amounts to. Thanks to Oliver Sensen for discussion. These issues are complicated partly by the fact that we can define normatively insignificant senses of "normative reason," distinct from "genuinely and robustly normative reasons," as e.g. Copp does in "Toward a Pluralist and Teleological Theory of Normativity." In one sense, *any* norm *N* gives rise to "reasons"—*N*-relative reasons. I avoid these complications here: talk of normative reasons in the text always refers to genuinely and robustly normative reasons.

porary theorizing in the area.[52] *Reasons Centrism* underlies the point of formulating debates about the extent of morality's authority in terms of reasons for action.[53] Relatedly, *Reasons Centrism* is part of what makes subjectivist views, on which reasons depend on agents' contingent desires, *prima facie* troubling. For given *Reasons Centrism* and subjectivism, moral requirements have authority for *A* only if *A* has a suitable contingent desire, and in this sense only hypothetically. Without *Reasons Centrism*, it is unclear why it would make good sense to prosecute questions about morality's authority in terms of reasons for action, or to worry about subjectivism's implications for morality's authority, as is usually done.

The second, related theoretical role for reasons for action is their link to what one ought to do, in what is sometimes called a "robustly normative" sense. Unlike merely qualified oughts, such as ought-according-to-etiquette, or ought-according-to-the-laws-of-England, it is usually assumed that robustly normative oughts, just as such, bear an important link to reasons.[54] Specifically, it is often assumed that, if *A* has decisive reasons to ϕ, then *A* ought in the robustly normative sense to ϕ. Further, it is important that such "oughts" can have the force of authoritative *demands* on agents to ϕ, not mere recommendations. We are often interested in whether we would be violating an important, authoritative demand on us if we failed to ϕ (e.g., "*Must* I do what morality requires in this instance, given the cost to myself?"), not just in whether there is some normative support for ϕ-ing, though it is still entirely normatively optional to ϕ. Of course, perhaps *not all* reasons impose demands: perhaps some merely entice or recommend (enticing reasons), or merely justify doing something that one would otherwise have been required to avoid (justifying reasons).[55] Still, the idea that (some) reasons can impose authoritative demands surely accounts for much of the theoretical interest in the notion of reasons for action. Merely "justifying" reasons would not be of much interest on their own, if there were no potential authoritative demands for them to oppose or disarm. And merely "enticing" rea-

52 One might still deny it, of course. Some of Kate Manne's work ("Internalism about Reasons" 113–16) may push in this direction.

53 It also underlies the point of asking "constitutivists" about practical reasons what reason we have to abide by agency's constitutive standards: this is a way of asking about these standards' authority or normative importance (Enoch, "Agency, Shmagency").

54 See, e.g., Broome, *Rationality through Reasoning*; McPherson, "Against Quietist Normative Realism." I remain neutral on whether "ought" in English is semantically ambiguous, or univocal but picks out different norms in different contexts of use.

55 See Gert, "Requiring and Justifying"; Portmore, *Commonsense Consequentialism*, ch. 5; Dancy, *Ethics without Principles*, 21.

sons are not very important as normative phenomena, since even the strongest and unopposed enticing reasons to ϕ leave ϕ-ing entirely normatively optional.

In sum, part of the distinctive point of appeals to reasons for action is captured by *Reasons Centrism*, along with the related idea that decisive reasons can impose authoritative demands. My argument for RR will be that, unless reasons are linked to good deliberation à la RR, they cannot impose authoritative demands on the agents whose reasons they putatively are. RR is a condition of reasons' playing the roles indicated. Sections 4–5 develop this argument and respond to objections.

4. REASONS, DELIBERATION, AUTHORITATIVE DEMANDS, AND REASONABLE EXPECTATIONS

Suppose RR is false.[56] Then there might be decisive reasons for *A* to ϕ in a situation, imposing an authoritative demand on *A* to ϕ, although there is no good deliberative route that A could take in that situation that would lead her from considering the reason-giving and other relevant facts to ϕ-ing. For instance, Sullenberger might be under an authoritative demand to steer left, given facts about the plane's behavior and what it takes to land safely, but there might be no possible rational response to these facts that would lead Sullenberger to steer left on their basis. Sullenberger might still end up steering left, of course. Perhaps he happens to do so for no (motivating) reason, or via some lucky nonrational response to considering the reason-giving facts (perhaps considering them causes his arms to twitch so as to steer left at the ideal moment). Or perhaps he is somehow led to steer left via considering some *other* facts entirely—say, about today's crossword puzzle—that do nothing to favor this maneuver. Still, if we deny RR, we deny that there is a recognizably rational response to the reason-giving facts of the situation that Sullenberger could have, that would take him from considering those facts to doing what they demand.

This is certainly *prima facie* odd.[57] We usually think there is a point to trying to figure out facts relevant to what we should do, and to trying to make important decisions in their light, instead of ignoring them. Likewise, we usually assume we will make better decisions if we consider the reason-giving facts ra-

56 The argument of this section draws in part on earlier attempts to articulate essentially the same argument in Paakkunainen, "Internalism and Externalism about Reasons" and "Normativity and Agency."

57 Cf. Schroeder's remark that it would be "puzzling to think that correct deliberation from complete information could lead us astray from what we ought to do" (*Slaves of the Passions*, 132).

tionally; thus we try to ensure we are not intoxicated, exhausted, or otherwise incapacitated when we have to make important decisions. In sum, we often seem to tacitly assume that deliberating well in light of the reason-giving facts helps us to do what we ought. Still, what exactly rules out the *prima facie* odd possibility that, at least sometimes, there is simply no way for us to rationally respond to the reason-giving facts by taking account of and acting on them, in the way they demand?

What rules it out is a further conceptual connection between authoritative demands and *reasonable expectations*. If the facts in a circumstance impose an authoritative demand on *A* to ϕ, then it must be reasonable to expect of *A* that she ϕ, at least under some conditions. More carefully, if some facts, *p*, demand of *A* that she ϕ in *C*, then there must be some possible condition *X* that *A* might be in in *C*, such that if *A* encounters *C* while in condition *X*, then it is reasonable to expect of *A* that she will ϕ. The operative notion of reasonable expectations is partly predictive, partly normative.

The predictive part is that, if *A* is in condition *X* in *C*, then *A* will ϕ in *C*; moreover, it is predictable that the action, ϕ, that *X* leads *A* to do is precisely the reason-demanded action. That is, there is a predictable match between what the reasons, *p*, demand of *A*, and what *X* leads *A* to do: condition *X well-equips A* to do, in *C*, just what the reasons demand. Unless there is some such possible condition *X*, the idea that the reasons can authoritatively demand *A* to ϕ seems to lapse. Again, *A* might of course happen, by sheer accident, to do precisely ϕ. And we might think of this accident as a happy one, applying some positive evaluative predicate to it.[58] But ϕ-ing is not something that reasons can authoritatively demand of *A*, if there is no possible condition, *X*, that well-equips *A* to meet those demands. So it seems to me.

What kind of condition might *X* be, then? Here the normative dimension of the operative notion of reasonable expectations is relevant. Reasons cannot impose demands on *A* to ϕ that it is completely unreasonable to expect *of A* that she fulfill. Reasons *can* sometimes demand *difficult* things of us—whether psychologically or physically difficult. But crucially, failures to do the reason-de-

58 Compare Railton, "How Thinking about Character and Utilitarianism Might Lead to Rethinking the Character of Utilitarianism," 408–10, on "morally fortunate" actions. Railton proposes "valoric utilitarianism" as a theory of "moral rightness" conceived of as classifying actions as morally fortunate or unfortunate. Rightness thus conceived may lack commonsense links between rightness and obligation on the one hand, and blameworthiness and what can be reasonably expected of people on the other. Railton acknowledges that such a theory risks being a mere classificatory scheme, and needs to explain why its notion of rightness still answers to concerns that make moral rightness an important notion to theorize in the first place.

manded thing, whether difficult or not, must be *apt targets of criticism*—at least absent excusing conditions.[59] Unless such failures are apt targets of criticism, the putative demand or expectation to ϕ lapses as wholly unreasonable. This dimension of apt criticizability will help us to determine the kind of condition that the well-equipping condition, *X*, must be.

When are putative demands wholly unreasonable and so not really in place, and when are they reasonable but such that A is excused for failing to meet them? Plausibly, nonculpable ignorance of relevant nonnormative facts, such as the reason-giving facts, *p*, excuses, while leaving the reasons and associated demands in place.[60] It is controversial what further conditions excuse. Extreme physical or psychological difficulty in meeting a demand plausibly excuses, at least sometimes. (Suppose one has post-traumatic stress disorder, or was drugged against one's will. On the other hand, when mere viciousness makes doing good things hard, perhaps one is not excused). Perhaps, at the limit, physical or psychological *impossibility* makes would-be demands lapse entirely: plausibly, reasons simply cannot demand us to fly unaided to the moon, for example.[61] I return to psychological impossibility below. For now, focus on the claim, above, that when there is no condition, *X*, that *A* could (metaphysically) be in in *C*, that would well-equip *A* to do the reason-demanded thing in *C*, then the putative demand lapses. This still seems true. And criticism of *A* plainly is not apt in such cases. But suppose there *is* such a well-equipping condition, *X*, that *A* could be in. What kind of a condition must *X* be, specifically, such that its availability helps to make sense of the aptness of criticism (absent excuses), should *A* still fail to do what the reasons demand?

It seems that *X* must, at least, be some *deliberative* condition of *A*'s. Consider: just as rocks or bugs are not subject to authoritative demands to behave in certain ways, nor are they aptly criticizable for failures to behave in those ways—even though there are conditions of rocks or bugs that well-equip them for certain behaviors in certain circumstances. (We can *evaluate* rocks as good or bad *qua* doorstops, say, but it makes little sense to demand them to perform better in

59 Not that some actual person must be able to appropriately level criticism, or some type of angry blame, at A if she fails to ϕ; rather, criticism for failures to ϕ is in principle appropriate. Thanks to a referee for prompting me to clarify this.

60 As before, the relevant reasons are thus objective reasons in the sense of not being relative to *A*'s information state (with some complications; cf. note 24 above). Do mistaken *normative* beliefs excuse? See Harman, "The Irrelevance of Moral Uncertainty," for illuminating discussion and a negative answer, regarding mistaken moral beliefs and excuses from moral blame in particular.

61 The example is from Sinclair, "Promotionalism, Motivationalism, and Reasons to Perform Physically Impossible Actions," 650–51.

this regard, or to criticize them if they do not.) Likewise, very young infants are not aptly criticizable for failures to behave in certain ways, any more than they are subject to authoritative demands to behave in those ways. Nor are we, otherwise mature, thinking agents, aptly criticizable for having or failing to have involuntary, nonrational twitches. What we may be criticizable for is having done or failed to do, as an exercise of agency, something to cause such twitches. More generally, apt criticism attendant to failures to do the reason-demanded thing, ϕ, concerns failures to ϕ *under one's own steam, as an exercise of agency*. Hence it seems that the well-equipping condition, *X*, whose availability helps to make sense of the aptness of criticism for failures to ϕ, must be some broadly deliberative condition of *A*'s: a condition in which *A* acts via what is at least a minimally rational, as opposed to *non*rational, response to features of the circumstance. Finally, though, *X* cannot be just *any* old deliberative condition. Since *X* must well-equip *A* to do precisely the reason-demanded action in *C*, *X* must somehow make *A* sensitive to the reasons in *C*. Absent such a possible deliberative condition, the putative demands, along with the aptness of criticizing *A* for failures to meet these demands, lapses.

So far I have argued that if some reasons, *p*, authoritatively demand of *A* that she ϕ in *C*, then there must be some (a) deliberative condition, *X*, that *A* could be in in *C*, that (b) well-equips *A* to do, in *C*, precisely the reason-demanded action, by making *A* somehow sensitive to the reasons in *C*. The question that remains is what deliberative condition *X* best meets condition (b).

The obvious answer is: the condition of deliberating well, in light of all the reason-giving facts (along with enablers and disablers), where good deliberation in light of these facts involves weighing them correctly and being moved to do what the weightiest reasons support. It is certainly hard to see how deliberating *badly*—giving some facts inadequate or disproportionate weight in deliberation, or drawing bizarre conclusions from them—might well-equip, or better-equip, agents to do just what the facts in a situation demand of them. Likewise, it is hard to see how rational responses to *misleading or irrelevant information*, information that does nothing to support ϕ-ing, would generally lead agents to ϕ.[62] Such responses would seem, precisely, *not* to be sensitive to the reasons in the circumstance so as to well-equip *A* to ϕ. This is the key reason why RR is so plausible. RR is so plausible because responding to the reason-giving facts by weighing them accurately, and by being moved by the weightiest reasons to do what they support, is the best candidate deliberative condition to well-equip agents to meet the demands imposed by their reasons.

62 However, I consider an objection in this vein in section 5.

I consider a major objection in section 5. For now, three important clarifications.

First, there may be stronger conditions on reasons than RR. I argued that, if reasons impose authoritative demands on *A* to ϕ, there must *at least* be a deliberative condition that *A metaphysically could* be in, that would well-equip *A* to meet those demands. Absent some such possible condition of *A*'s, the putative demands on *A* lapse. But while the relevant metaphysical possibility is necessary to ensure that reasons' demands do not lapse, perhaps it is insufficient. Specifically, if it is *psychologically impossible* for *A* to be in the relevant deliberative condition, perhaps the reasons still lapse; correlatively, perhaps criticism given such psychological impossibility is inapt, not merely something that *A* is excused from. Further, perhaps mere psychological possibility of being in the relevant deliberative condition is not enough to ensure that the reasons do not lapse, either. If the psychologically possible worlds in which *A* is in the condition are very remote, then again it may seem dubious whether the reasons remain in force, and whether criticism for failures to abide by them is apt. (Compare criticizing someone for a failure to abide by a putative demand to swim across the English Channel, in pursuit of a large reward for a lifesaving charity, where a successful swim, while physically possible, is vanishingly unlikely.) This might motivate an even stronger, "internalist" constraint on reasons, on which *A*'s reasons to ϕ depend on *A*'s having the present actual *motivational capacity* to ϕ via rational sensitivity to those reasons. Having a present actual capacity to ϕ (e.g., swim, deliberate well) implies not just that one ϕs in some possibly faraway world, but roughly, that there is a robust range of relatively nearby worlds in which one ϕs (if one tries).[63]

Nothing I have said rules out such stricter constraints on reasons. Whether my argument can be positively extended to support them is a complex question I cannot adjudicate here. Still, if I am right, adjudicating it involves determining what kinds of impediment to exercising rational sensitivity to reasons merely excuse from otherwise reasonable criticism, leaving the reasons and their demands in place, and which impediments make the reasons lapse entirely. I doubt that there is a very neat way of deciding these questions. For instance, whether lack of present motivational capacity to ϕ via rational sensitivity to the reasons for ϕ-ing *even excuses* failures to ϕ may depend on the cause of the lack. Vicious

63 This is inexact, but suffices here. I consider motivational capacity claims, and the relevant kind of internalism, in more detail in Paakkunainen, "Internalism and Externalism about Reasons." Thanks to a referee and Benjamin Wald for prompting me to clarify my stance about these issues here. Cf. Lord, "Acting for the Right Reasons, Abilities, and Obligation," who argues for an epistemic and an ability condition on reasons—conditions that are also stronger than RR—on grounds somewhat similar to those I use to support RR.

people may lack present motivational capacities to be moved by reasons to act well, but if they ought to have acquired such capacities and voluntarily did not, then perhaps the lack does not even excuse.[64] All the same, my argument for RR stands. Whether or not stronger constraints can be supported on similar grounds, *at the very least*, RR is a condition of reasons' imposing authoritative demands. If *A* metaphysically could not be in a deliberative condition that well-equips her to do the reason-demanded thing, ϕ, the putative demand on *A* to ϕ lapses. And the best candidate well-equipping deliberative condition, I argued, is that of deliberating well *tout court* à la RR.

The second clarification concerns the operative notions of authoritative demands, reasonable expectations, and apt criticism. We might worry that these notions are too obscure to guide our judgments in the area. For example, it is hard to precisely characterize the kind of criticism connected to failures to meet authoritative demands. Not *all* kinds of criticism are so connected: we sometimes seem to make "personal" criticisms of people for, e.g., having poor taste or for being boring, where such criticisms *need not* imply that their targets are under authoritative demands to be different, and likewise *do not* turn on the availability of some possible deliberative condition such that, if the agents were in it, they would be well-equipped to be less boring or to have better taste. But how can we assess my argument for RR without a better grip on the relevant notions and when they apply? Do we not first need a clear account of what it is for someone to be under authoritative demands, and of the attendant sort of criticism, in order to judge the merits of my argument?[65]

In response, I doubt we could first adequately characterize the operative notions, completely independent of their links to reasons and good deliberation, and then use them as precise criteria for assessing my argument for RR. If my argument works, it is because there is a cluster of concepts here that hang together, and whose interrelations help us, via reflection of the sort I engaged in, to clarify the contours of these very concepts—including the concept of a reason for action. Further, we can admit that criticism for *other* sorts of failures, apart from failures to do the reason-demanded thing, may be apt even *absent* the metaphysical possibility of avoiding the failure through good deliberation. Perhaps personal criticisms for being boring or having bad taste are like this; I am not sure. Still, as I argued, criticism attendant to failures for doing a reason-demanded action *does* seem to require, for its aptness, the relevant deliberative possibility. Like-

64 Thanks to Tristram McPherson for discussion.

65 Thanks to a referee for articulating this challenge.

wise, unless criticism *is* apt for failures to do the reason-demanded thing (absent excuses), the putative demands seem to lapse.[66]

I cannot say much more than I have said to support the conceptual connections I have been tracing. Still, reflection on a different kind of normative fact concerning agent *A*, one that *does not* involve an authoritative demand on *A*, may help those skeptical of these connections. Suppose Alex ought to be amply rewarded. It is not that she ought to secure the reward for herself, and violates an authoritative demand on her if she does not. Rather, someone else ought to reward her. In such a case, it clearly does not follow that it is reasonable to expect *of Alex* that she be rewarded. It only seems reasonable to expect it of the person who ought to reward Alex. And if Alex is not rewarded, it seems inapt to criticize *her* for that; again, criticism is only apt for the person who ought to reward Alex. Likewise, there need be no possible deliberative condition *of Alex's* that would well-equip her to secure the reward in the situation. Perhaps it is simply not up to Alex whether she gets rewarded, no matter how well she deliberates. Finally, whatever makes the claim "Alex ought to be amply rewarded" true (in context), it seems clear that it is not something about Alex's (present) normative reasons for action. (Though it may involve Alex's having done something she previously had good reason to do.)[67]

So it seems that ought-facts concerning *A* that *do not* imply an authoritative demand on *A* likewise *are not* linked to what *A* could be reasonably expected to do, would be aptly criticizable for, or could do through a course of good deliberation; nor are they linked to *A*'s reasons for action. I think it is no accident that these verdicts cluster in this way—as they seemed to cluster in the positive case where *A*'s reasons *do* impose authoritative demands on her. These appearances are some further confirmation that the links drawn between reasons, authoritative demands, reasonable expectations, apt criticism, and good deliberation trace important joints in our normative concepts.

Finally, the third clarification concerns the requirement that, where A's reasons demand her to ϕ, it must be that *A herself* could deliberate to the conclusion to ϕ, via sensitivity to her reasons. Why is it not enough if someone else—a trusted advisor—deliberates on *A*'s behalf, telling *A* that she ought to ϕ, though

66 A slightly different challenge in the vicinity is that we need at least *some sense* of the operative notions of authoritative demands and apt criticism, even if not a clear account, to judge the merits of my argument. But again, the reflections in this section trade on, as well as serve to clarify, such a sense.

67 The example is adapted from Broome, *Rationality through Reasoning*, 65.

keeping the reasons why hidden from *A*? Better still, perhaps the mere *possibility* of such a trusted advisor, and of shared deliberative labor, is enough?[68]

Here is why this does not work. Suppose the advisor is actually present, but gets it wrong, telling *A* to ψ instead of to ϕ. If it is not even metaphysically possible for *A* herself to better deliberate to the right conclusion, then it seems that, again, *A* cannot be aptly criticized for failing to ϕ, and the putative demand on *A* to ϕ seems to lapse. (This is perhaps clearest where *A* had decisive reasons to trust the advisor.) The mere possibility of an advisor who gets it right does not seem to change this verdict. But what if the actual advisor gets it right? Perhaps at least in *such* cases it need not be that *A* could herself deliberate well to the conclusion to ϕ? But it would be odd for the presence or absence of *A*'s reasons to ϕ to depend on the serendipitous presence of an actual advisor who gets her advice right. Note that the converse judgment does not seem plausible: if it were for some reason metaphysically impossible for *A* to have a trusted advisor of the relevant sort, but *A* herself had (perhaps serendipitously) a generally excellent, though not infallible, sensitivity to reasons, we would not *thereby* think that reasons cannot impose demands on *A* (even where *A* in fact gets it wrong).

Shared deliberative labor can happen, and is sometimes useful. Outsourcing deliberative labor to experts often minimizes our risk of going wrong. Still, if we metaphysically could not have gotten it right ourselves, and the experts lead us astray, reasons cannot demand us to ϕ. And the presence of reasons for us to ϕ does not depend on whether an actual expert accurately tracks them on our behalf.[69]

5. INDIRECTION RECONSIDERED

Suppose I am right that, if reasons demand ϕ-ing of *A* in a situation, then there is a possible deliberative condition of *A*'s that well-equips *A* to do precisely the demanded action in that situation. We might nonetheless doubt my further claim that the best candidate deliberative condition is that of deliberating well *tout court*—responding rationally to the reason-giving and other relevant facts, by weighing them accurately and being moved by the weightiest reasons to do what they demand. We might doubt this claim for reasons similar to those fueling government house or "indirect" utilitarianism.

68 Thanks to a referee for raising this concern. Cf. Sinclair, "On the Connection between Normative Reasons and the Possibility of Acting for those Reasons."

69 A more complete treatment would discuss cases such as Sinclair's Tate/Loco case ("On the Connection between Normative Reasons and the Possibility of Acting for those Reasons," 1220), but I cannot do that here.

Recall the key thought behind indirect utilitarianism: while agents morally ought to maximize utility, they are more likely to do so if their deliberations ignore considerations about impartial utility-maximization, and focus on ordinary considerations of, e.g., friendship, personal loyalty, or love. For utility calculations are hard and often take too long for actual agents to complete before action is required. Further, always focusing on impartial utility considerations would deprive agents of personal attachments that themselves have significant utility. One cannot really love a person if this person and their specific concerns never weighs more heavily in one's deliberations than the concerns of strangers. Finally, and related: even if impartial utility calculations allowed agents to weight their loved ones' interests more heavily than the interests of strangers—perhaps because special bonds between people contribute heavily to utility, and each agent is best positioned to maintain their own special bonds—we might still worry that, if loved ones' concerns are given special weight in deliberation only *because* doing so is utility-maximizing, or *under the guise* of utility-maximization, this again is incompatible with real love, the very thing whose contribution to utility we are concerned to preserve.[70]

We might raise similar concerns about deliberating in terms of the reason-giving facts.[71] First, taking account of all the reason-giving facts—even the important ones—is sometimes too hard or takes too long. Perhaps shortcuts or rule-of-thumb policies of deliberation often better-equip actual agents to do what their reasons demand than does attempting to take account of the reason-giving facts à la RR. (Perhaps such policies provide a kind of indirect sensitivity to reasons' demands.) Second, agents often have reasons to act in loving, kind, loyal, modest, etc., ways; indeed, plausibly the very fact that ϕ-ing would be loving, kind, etc., is *itself* a reason, and an important one, to ϕ. After all, being loving, kind, etc., are important good-making features of acts, and plausibly, important good-making features of acts provide reasons to perform those acts. Yet (so the objection goes) ϕ-ing is an act of love, kindness, etc., only if, in ϕ-ing, one *does not* act on the consideration that ϕ-ing is an act of love or kindness. Acts of love or kindness are motivated by direct concerns for loved ones' interests, or for others' needs: concerns not mediated by a further concern with whether the actions would express positive characteristics in oneself. In short, the second concern is that, while the fact that ϕ-ing would be an act of (say) love or kindness is a reason to

70 For classic discussion, see, e.g., Williams, *Ethics and the Limits of Philosophy*, 107–10; Railton, "Alienation, Consequentialism, and the Demands of Morality" and "How Thinking about Character and Utilitarianism Might Lead to Rethinking the Character of Utilitarianism."

71 I will not worry about keeping the concerns below exactly parallel to those raised above. I will simply try to raise the best objections in the vicinity.

ϕ, this fact, and so the reason, would be destroyed if it were among the considerations that move one to act.[72]

I address the two concerns in order. The second one motivates small qualifications to RR, but these are independently well motivated, and preserve the core thought in RR and my argument for it.

As noted in sections 1–2, imperfect agents may do well to rely on rule-of-thumb policies when they are unlikely to respond rationally to considering the reason-giving facts, perhaps due to getting unnerved. This does not conflict with RR. However, the present objection *is not* that the usefulness of rule-of-thumb policies conflicts with RR. It is rather that the *argument* for RR in section 4 misfires, given the usefulness of such policies. That argument claimed that responding rationally to the reason-giving facts—via direct sensitivity to them, à la RR—is the best candidate deliberative condition to well-equip *A* to meet her reasons' demands. The present objection is that it is not: often the best deliberative condition to well-equip agents to do reason-demanded actions is that of applying rule-of-thumb policies that ignore the reason-giving facts. Such policies often better lead imperfect agents to doing what they ought than does attempting to deliberate well *tout court*, in ways they are unlikely to succeed in doing.

This objection is ultimately confused. The argument of section 4 turns on identifying the deliberative condition that would best equip agents to do what their reasons demand in a situation, were they to be in that condition. The condition of deliberating well *tout court*, where this involves weighing reasons accurately and being moved by the weightiest reasons to do what they support, still seems like the best candidate. Following rules of thumb may indeed often better lead imperfect agents to doing what they ought than would *attempting* to deliberate well *tout court*. But if so, this is because these agents' imperfections make them unlikely to deliberate well *tout court* in the situation at hand. This is compatible with holding that, *were* these agents to deliberate well *tout court* in their situation, they would be even better equipped to meet the demands of the situation.[73]

72 Thanks to Massimo Renzo and a referee for formulating versions of these objections especially clearly. The second objection resembles *Surprise Party* or *James Bond*, except that, in those cases, mere *belief* that *p* is supposed to destroy *p*'s status as a reason (perhaps by destroying *p*'s status as a fact). In the present objection, *A* might perhaps believe that ϕ-ing would be an act of (say) love, and still perform ϕ-ing out of love. But the fact that ϕ-ing would be an act of love allegedly cannot be among the considerations on which one acts, without destroying the act's status as an act of love, thereby destroying the reason.

73 The assumption that A deliberates well *tout court* must not destroy *the imperfections on which A's reasons depend*; for it would thereby destroy the supposed reasons. But disappearing the imperfections that prevent *A* from undertaking a particular course of good deliberation

This point may seem merely nitpicky, but its force is evident where rules of thumb lead astray. Indirect utilitarians readily admit that rules of thumb can lead astray. Deliberating on the basis of, say, considerations of love or loyalty may often maximize utility, but sometimes it will not.[74] Similar scenarios can arise for reasons and the rule-of-thumb policies of deliberation that ignore them—as with the soldier who ignores the common humanity of the enemy inhabitants even when it is relevant to which combat maneuvers to perform. Where rules of thumb do lead astray, following them plainly is not the deliberative condition that best equips you to do what you ought. If you follow a rule of thumb accurately when it leads you astray, you precisely will not do what you ought. You might still accidentally end up doing what you ought if you follow the rule of thumb badly, or if you do not deliberate at all and simply act on the basis of no considerations whatsoever. But neither of these latter strategies plausibly better equips you to do what you ought than does deliberating well *tout court*.

Hence following rule-of-thumb policies cannot be the best candidate deliberative condition, across the board, to well-equip agents to meet their reasons' demands. The best candidate deliberative condition still seems to be deliberating well *tout court*. In response, one might propose a disjunctive condition on reasons: there must be *either* a rule-of-thumb policy that ignores one's reasons, *or* a deliberative route that takes them into account à la RR. This would leave room for GH reasons. But, as always, a more unified view is preferable if it has no significant costs. So far RR seems to have no significant costs: neither *Surprise Party*–style cases nor *Soldier*-type cases turned up such costs (section 2). We should prefer the simple and natural view that reasons relate to good deliberation directly, as premises to be weighed and acted upon, to disjunctive views that have no clear benefit.

However, perhaps the second concern above, regarding acts of love, kindness, etc., motivates a disjunctive view. Facts such as that ϕ-ing would be kind or loving do seem to be reasons—indeed, often decisive ones—to ϕ. Denying

need not destroy the imperfections on which the relevant reasons depend. Take the famous "sore loser" example (Smith, "Internal Reasons," 111–12; Watson, "Free Agency"): *A*'s intemperance gives her reasons to avoid shaking her winning opponent's hand, reasons that a more temperate version of *A* would lack. Still, there can be a course of good deliberation that takes into account *A*'s actual intemperate self's reasons in the situation, weighs them accurately and leads to *A*'s being moved by the weightiest reasons to do what they support. This course of deliberation, like the reasons it registers, differs from the course of good deliberation that *A*'s more temperate counterpart would engage in.

74 See, e.g., Railton, "Alienation, Consequentialism, and the Demands of Morality," 157, and "How Thinking about Character and Utilitarianism Might Lead to Rethinking the Character of Utilitarianism," 402.

that they are reasons seems to carry a higher intuitive cost than denying that, e.g., the specific fact that a surprise party awaits—instead of some other, nearby fact, such as that Nate would be glad if he went home—is a reason for Nate go home. Yet it seems that deliberating well *tout court* is not an option in these cases, compatibly with the reasons remaining on the scene. This may be the strongest intuitive challenge to RR, in the vicinity of *Surprise Party*–style cases.

I will make two main points in response. First, it seems that one *can* act in, e.g., a loving way while basing one's act, at least partly, on the fact that the act would be loving. We often try to figure out what "the loving thing to do" would be, whether in our own case or when advising others. And when an advisor gives advice in this regard, we would not naturally expect them to immediately also advise us to forget the advice. We understand an action's being the loving thing to do as a significant reason in its favor, and one that is not to be ignored. Admittedly, it is doubtful that an act would really be loving if its being loving were the *only* consideration the agent took into account. It seems that those performing genuinely loving acts also have some grasp of *why* the act would be loving, and they act partly on these further considerations. Still, it seems that one can act lovingly while having "this is the loving thing to do" figure as a summary judgment in one's deliberations, summing up the kind of importance that various other facts of the case, the facts that make the act a loving one, have.

Likewise for kindness, and even modesty. It seems that actions can be modest while being based, partly, on the fact that this is "the modest thing to do," or that "doing otherwise would be immodest." Recall that taking such considerations into account can be swift, as it plausibly is in genuinely modest people. Taking them into account does not require self-congratulatory lingering on thoughts about how modest one is about to be. It is such self-congratulatory lingering that seems modesty-undermining in these cases, not the swift registering of the fact that this is the modest thing to do. Deliberation preceding modest, loving, or kind acts plausibly spends most time focused on other considerations. Nor will modest, loving, or kind agents try to draw others' attention to the good qualities of themselves or their actions.[75] But this is compatible with basing one's action partly on the consideration that it is the modest, kind, or loving thing to do.

This defuses much of the objection involving loving, kind, modest, etc., actions, but not all of it. We previously claimed that good deliberation takes account of *all* the important reasons, lest it lead agents astray from what the balance of reasons supports (section 1). Yet one can act in loving, kind, modest

75 Compare Star's helpful discussion of modesty, on which I partly draw; though I am unsure he would grant that acting on the consideration that "this is the modest thing to do" is compatible with the act's modesty (*Knowing Better*, 83–85).

ways *without* considering the act's status as loving, kind, or modest; indeed, the paradigm cases of loving, kind, modest actions are like this. And surely deliberation that ignores the fact that the action would be (say) loving is not faulty for that—and need not risk leading the loving agent astray. If so, and if the fact that ϕ-ing would be loving is an important reason to ϕ, then RR must be qualified.

This brings me to the second main point of response. When an act is, say, loving or kind, further features of the act make it so. Perhaps the act serves a loved one's or a stranger's needs, say. These further features seem to *also* provide reasons to perform the relevant acts. But crucially, the two sets of considerations—one about others' needs, the other about the act's being loving or kind—do not constitute two independent normative cases in favor of the act. Consider: if acting lovingly or kindly somehow involved trampling others' needs, then an act's being loving or kind would not seem like much of a mark in its favor. Much of the normative import of an act's being loving or kind seems bound up with the normative import of others' needs.[76] "Bound up" how? I will describe two options, independently well motivated, each of which seems to be actualized in some cases. We need not decide which option better fits cases of loving/kind actions. The key point is that each option fits a slightly modified RR, while preserving the spirit of the argument for RR.

Option 1: Reasons are not individual, fine-grained facts, but clusters of facts. Context determines which facts in the cluster it is most natural for agents to consider in deliberation. (Or to give as advice, or as explanations of actions after deliberation.) Often when one fact is identified as a reason for *A* to ϕ, many nearby facts look like equally good candidate reasons for *A* to ϕ—indeed, reasons of equal strength. Yet we should not count all of these facts as reasons separately: that would yield a cumulative case for ϕ-ing that is too strong. If I ought to visit my mother soon, we might cite the decisive reason as the fact that she is sick, or that she has pneumonia, or that, being sick, she will need some care. These do not each add independent weight to the case for visiting my mother. Yet each is a good candidate reason, and it is unclear how to choose which "the" reason is. Further, it is a rational response to each of these facts for me to go visit my mother: each candidate reason corresponds to a good deliberative route. This phenomenon seems extremely common. This is why it is natural to view reasons as clusters of facts, where context determines—with some room for arbitrariness—which individual facts get called "the reasons" in the situation, and get to be the focus in deliberation.[77]

76 While this seems less plausible to me, we might also doubt the normative import of facts about others' needs if tending to them were somehow systematically unloving or unkind.

77 Conversations with Daniel Fogal led me to develop this suggestion. For related discussion,

Applying this idea to loving/kind actions, the fact that an act is (say) loving and the facts that make it loving belong to a cluster of facts that collectively call for specific actions. Depending on context, only some of these facts are the focus of a given course of good deliberation, or of advice. We can modify RR to accommodate these points. What is important is that, given decisive reasons to ϕ—whichever facts get called "the reasons" in the context—a good deliberative route takes you from considering these reasons to ϕ-ing. The core thought in the argument for RR remains: the best deliberative condition to well-equip agents to do what their reasons demand is to take account of key facts in the relevant cluster, weighting them accurately and doing what they (together with other facts in the cluster) demand. This still differs crucially from deliberation that focuses on non-reason-giving considerations; yet it explains why, with loving/kind actions, we can deliberate well *without* considering the specific fact that these actions are loving or kind.

Option 2 (not really in tension with the "cluster" idea but going beyond it): Some reasons are normatively *fundamental*, while others *derive* their normative import from the fundamental reasons. Suppose ϕ-ing would destroy the only crop-yielding field in a village, and destroying this field would lead to much suffering. Both facts seem to be reasons—indeed, decisive ones—against ϕ-ing. But the status of the first fact as a reason against ϕ-ing seems derivative from the status of second fact, about suffering, as a reason against ϕ-ing. The second fact is more fundamental as a reason, while the first fact is a derivative one: it is a reason *because* destroying the field would cause much suffering.[78] Here, a rational response to *either* the derivative *or* the (more) fundamental reasons would lead you to do what the balance of reasons supports. We should construe RR so that each counts as a good deliberative route. Where neither rational focus on solely the derivative reasons nor rational focus on solely the fundamental reasons would lead you astray from what you ought to do, good deliberation can safely focus on just one or another set of reasons. (And deliberation that *does* consider *both* sets of reasons should not count each set as independently weighty, on pain of overestimating the case against ϕ-ing.)

Applying this idea to loving/kind actions, some of the reasons in the case are derivative, others fundamental. (We need not decide which are which.) Where a rational response to *either* the fact that the act is loving/kind, *or* to the facts that *make* the act loving/kind, would lead to doing what the balance of reasons

see Raz, *From Normativity to Responsibility*, 16–17.

78 See Star, *Knowing Better*, for a thorough discussion of fundamental versus derivative reasons. Star argues that his theory of reasons to ϕ as evidence that one ought to ϕ accommodates this phenomenon well.

supports, you can safely focus on just one set of facts while ignoring the other. The core thought in RR is that there must be a possible rational response to the important reasons in a situation that would lead you from considering those reasons to doing what the reasons demand of you. This core thought is preserved even if we acknowledge that there is leeway, because of dependency relations between reasons, as to which of them good deliberation focuses on.

I have argued that we can allow that an act's being, say, loving or kind is an important reason in its favor. One can act lovingly or kindly while taking these reasons into account. One can also act lovingly or kindly *without* taking these reasons into account, but this does not threaten RR or my argument for it either—not once we qualify RR in certain independently well motivated ways. Finally, the usefulness of rule-of-thumb policies of deliberation does not undermine my argument either.[79] The most pressing objections inspired by indirect utilitarianism fail. RR and my argument for it stand.

6. CONCLUSION

A version of the Deliberative Constraint holds because without it, we cannot make sense of the peculiar normativity of reasons for action: of their ability, when decisive, to impose authoritative demands on agents. I defended RR, with some qualifications, as a good formulation of the Deliberative Constraint. I grant that a different reason-concept, one that abandons RR, might play *other* important theoretical roles, even if it cannot play the roles associated with reasons' peculiar normativity. Indeed, for all I have said, there may be important theoretical roles that reasons that obey RR cannot play. I am unaware of any such roles. Reasons' role in giving and receiving advice, for example, or in explaining normatively well-supported action, seem compatible with RR. Still, this is territory ripe for further exploration. My view is that we should be as clear as possible about which theoretical and related everyday roles various notions can and cannot play, and subsequently about where to draw principled lines between importantly different reason-concepts.[80]

Despite its modesty, RR does have implications for further theorizing. First, since the Advice Model of reasons is compatible with the falsity of RR, that model is either false (if unsupplemented by RR) or incomplete—at least, as a model

79 Note further that given the distinction between fundamental and derivative reasons, it is somewhat plausible that rule-of-thumb policies work, to the extent that they do, by taking account of (what are in most cases) derivative reasons, even if they ignore the fundamental ones. However, I will not pursue this here.

80 Thanks to David Plunkett and Daniel Star for helpful discussion here.

of reasons capable of imposing authoritative demands on the agents whose reasons they are.[81] Likewise for other views that do not incorporate a commitment to RR. Second, evaluative facts about good deliberation constrain facts about reasons, and we may find first-order normative insight, as well as metaethical insight about how reasons function, by studying good practical reasoning. Third, metaphysical facts about what deliberation as such can be like also constrain facts about reasons, via constraining facts about what good deliberation can be like. This makes the study of agency necessary for understanding reasons. Finally, theories that explain why RR holds are preferable to theories that do not. RR is a key *explanandum* for theorists of reasons, good reasoning, and agency. It should be explicitly in the foreground as a guide to theorizing.[82]

I have only addressed reasons for action. Does some analogue of the Deliberative Constraint apply to reasons for belief or emotion, and does my argument extend to these cases? While I doubt that a similar argument will extend to these cases, there may be different arguments for epistemic or emotional analogues of the Constraint.[83] This is a further area ripe for exploration.[84]

Syracuse University
hpaakkun@syr.edu

81 Cf. Smith, "Internal Reasons" and "Reasons with Rationalism after All"; and my introduction, above. Note that RR is importantly distinct from the "Example Model" that is Smith's foil in defending the Advice Model. Unlike the Example Model, RR treats agents' imperfections in a way designed not to eliminate reasons that depend on those very imperfections. See notes 24, 38, 46, and 73 above for some relevant remarks.

82 In the way that, e.g., Setiya (*Reasons without Rationalism* and "What Is a Reason to Act?") and Way ("Reasons as Premises of Good Reasoning") foreground their versions of the Deliberative Constraint.

83 See, e.g., Hieronymi's "The Wrong Kind of Reason" and "The Use of Reasons in Thought (and the Use of Earmarks in Arguments)" for an argument that thinking of reasons as premises in reasoning helps to distinguish right from wrong kinds of reasons for attitudes.

84 For extremely helpful feedback on previous versions of this material, I am grateful to the participants of the 2015 NYU Abu Dhabi Workshop on Normativity and Reasoning, the participants of the 2015 CEPA Seminar at the Murphy Institute at Tulane University, the participants of the 2015 St. Louis Annual Conference on Reasons and Rationality, referees for this and one other journal, and especially to Samuel Asarnow, Daniel Fogal, Kim Frost, André Gallois, Tristram McPherson, David Plunkett, Massimo Renzo, Mark Schroeder, Oliver Sensen, Neil Sinhababu, Daniel Star, Sergio Tenenbaum, Benjamin Wald, Jonathan Way, and Ralph Wedgwood. I am grateful to the Center for Ethics and Public Affairs at the Murphy Institute at Tulane for supporting part of the work for this paper.

REFERENCES

Ashwell, Lauren. "Superficial Dispositionalism." *Australasian Journal of Philosophy* 88, no. 4 (December 2010): 635–53.

Broome, John. *Rationality through Reasoning*. Oxford: Oxford University Press, 2013.

Chang, Ruth. "Introduction." In *Incommensurability, Incomparability, and Practical Reason*, edited by Ruth Chang, 1–34. Cambridge, MA: Harvard University Press, 1997.

Clarke, Randolph. "Opposing Powers." *Philosophical Studies* 149, no. 2 (June 2010): 153–60.

Copp, David. "Moral Naturalism and Three Grades of Normativity." In *Morality in a Natural World*, 249–83. Cambridge: Cambridge University Press, 2007.

———. "Toward a Pluralist and Teleological Theory of Normativity." *Philosophical Issues* 19, no. 1 (October 2009): 21–37.

Dancy, Jonathan. *Ethics without Principles*. Oxford: Oxford University Press, 2004.

———. "Nonnaturalism." In *Oxford Handbook of Ethical Theory*, edited by David Copp, 122–43. New York: Oxford University Press, 2006.

Darwall, Stephen. *Impartial Reason*. Ithaca, NY: Cornell University Press, 1983.

Enoch, David. "Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action." *Philosophical Review* 115, no. 2 (April 2006): 169–198.

———. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press, 2011.

Finlay, Stephen. "Recent Work on Normativity." *Analysis* 70, no. 2 (April 2010): 331–46.

FitzPatrick, William. "Robust Ethical Realism, Non-Naturalism, and Normativity." *Oxford Studies in Metaethics* 3 (2008): 159–206.

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

Gert, Joshua. "Requiring and Justifying: Two Dimensions of Normative Strength." *Erkenntnis* 59, no. 1 (July 2003): 5–36.

Handfield, Toby, and Alexander Bird. "Dispositions, Rules, and Finks." *Philosophical Studies* 140, no. 2 (August 2008): 285–98.

Harman, Elizabeth. "The Irrelevance of Moral Uncertainty." *Oxford Studies in Metaethics* 10 (2010): 53–79.

Hieronymi, Pamela. "The Use of Reasons in Thought (and the Use of Earmarks in Arguments)." *Ethics* 124, no. 1 (October 2013): 114–27.

———. "The Wrong Kind of Reason." *Journal of Philosophy* 102, no. 9 (September 2005): 437–57.

Horty, John F. *Reasons as Defaults*. Oxford: Oxford University Press, 2012.

Johnson, Robert N. "Internal Reasons and the Conditional Fallacy." *The Philosophical Quarterly* 49, no. 194 (January 1999): 53–71.

———. "Internal Reasons: Reply to Brady, Van Roojen, and Gert." *The Philosophical Quarterly* 53, no. 213 (October 2003): 573–80.

Korsgaard, Christine M. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press, 2009.

———. "Skepticism about Practical Reason." *The Journal of Philosophy* 83 no. 1 (January 1986): 5–25.

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

Lord, Errol. "Acting for the Right Reasons, Abilities, and Obligation." *Oxford Studies in Metaethics* 10 (2015): 26–52.

Lord, Errol, and David Plunkett. "Reasons Internalism." In *The Routledge Handbook of Metaethics*, edited by David Plunkett and Tristram McPherson. Abingdon: Routledge, 2017.

Manne, Kate. "Internalism about Reasons: Sad but True?" *Philosophical Studies* 167, no. 1 (January 2014): 89–117.

Markovits, Julia. *Moral Reason*. Oxford: Oxford University Press, 2014.

McDowell, John. "Might There Be External Reasons?" In *World, Mind, and Ethics*, edited by J. E. J. Altham and Ross Harrison, 68–85. Cambridge: Cambridge University Press, 1995.

McPherson, Tristram. "Against Quietist Normative Realism." *Philosophical Studies* 154, no. 2 (June 2011): 223–40.

Millgram, Elijah. "Williams' Argument Against External Reasons." *Noûs* 30, no. 2 (June 1996): 197–220.

Paakkunainen, Hille. "Internalism and Externalism about Reasons." In *The Oxford Handbook of Reasons and Normativity*, edited by Daniel Star. Oxford: Oxford University Press, forthcoming.

———. "Normativity and Agency." In *The Routledge Handbook of Metaethics*, edited by David Plunkett and Tristram McPherson. Abingdon: Routledge, 2017.

Parfit, Derek. *On What Matters*. 2 vols. Oxford: Oxford University Press, 2011.

Portmore, Douglas W. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford: Oxford University Press, 2011.

Railton, Peter. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13, no. 2 (Spring 1984): 134–71.

———. "How Thinking about Character and Utilitarianism Might Lead to Re-

thinking the Character of Utilitarianism." *Midwest Studies in Philosophy* 13, no. 1 (September 1988): 398–416.

Raz, Joseph. *From Normativity to Responsibility*. Oxford: Oxford University Press, 2011.

———, ed. *Practical Reasoning*. Oxford: Oxford University Press, 1978.

———. "The Truth in Particularism." In *Engaging Reason: On the Theory of Value and Action,* 218–46. Oxford: Oxford University Press, 1999.

Scanlon, Thomas M. *Being Realistic about Reasons*. Oxford: Oxford University Press, 2014.

Schroeder, Mark. "Realism and Reduction: The Quest for Robustness." *Philosophers' Imprint* 5, no. 1 (February 2005): 1–18.

———. *Slaves of the Passions*. Oxford: Oxford University Press, 2007.

Setiya, Kieran. "Introduction: Internal Reasons." In *Internal Reasons: Contemporary Readings*, edited by Kieran Setiya and Hille Paakkunainen. Cambridge, MA: MIT Press, 2012.

———. *Reasons without Rationalism*. Princeton, NJ: Princeton University Press, 2007.

———. "Reply to Bratman and Smith." *Analysis* 69, no. 3 (July 2009): 531–40.

———. "What Is a Reason to Act?" *Philosophical Studies* 167, no. 2 (January 2014): 221–35.

Setiya, Kieran, and Hille Paakkunainen, eds. *Internal Reasons: Contemporary Readings*. Cambridge, MA: MIT Press, 2012.

Shafer-Landau, Russ. "A Defence of Categorical Reasons." *Proceedings of the Aristotelian Society* 109, no. 1 (August 2009): 189–206.

Sinclair, Neil. "On the Connection between Normative Reasons and the Possibility of Acting for those Reasons." *Ethical Theory and Moral Practice* 19, no. 5 (November 2016): 1211–23.

———. "Promotionalism, Motivationalism, and Reasons to Perform Physically Impossible Actions." *Ethical Theory and Moral Practice* 15, no. 5 (November 2012): 647–59.

Skorupski, John. *The Domain of Reasons*. Oxford: Oxford University Press, 2010.

Smith, Michael. "Internal Reasons." *Philosophy and Phenomenological Research* 55, no. 1 (March 1995): 109–31.

———. *The Moral Problem*. Malden, MA: Blackwell Publishing, 1994.

———. "Reasons with Rationalism After All." *Analysis* 69, no. 3 (July 2009): 521–30.

Sobel, David. "Explanation, Internalism, and Reasons for Action." *Social Philosophy and Policy* 18, no. 2 (Summer 2001): 218–35.

———. Review of *Slaves of the Passions*, by Mark Schroeder. *Notre Dame Philosophical Reviews* (2009). http://ndpr.nd.edu/news/slaves-of-the-passions.

Star, Daniel. *Knowing Better: Virtue, Deliberation, and Normative Ethics*. Oxford: Oxford University Press, 2015.

Sturgeon, Nicholas L. "Ethical Naturalism." In *Oxford Handbook of Ethical Theory*, edited by David Copp, 91–121. New York: Oxford University Press, 2006.

Wallace, R. Jay. "Constructivism about Normativity: Some Pitfalls." In *Constructivism in Practical Philosophy*, edited by James Lenman and Yonatan Shemmer, 18–39. Oxford: Oxford University Press, 2012.

Watson, Gary. "Free Agency." *The Journal of Philosophy* 72, no. 8 (April 1975): 205–20.

Way, Jonathan. "Reasons as Premises of Good Reasoning." *Pacific Philosophical Quarterly* 98, no. 2 (June 2017): 1–20.

Williams, Bernard. *Ethics and the Limits of Philosophy*. London: Fontana Press, 1985.

———. "Internal and External Reasons." In *Moral Luck*, 20–39. Cambridge: Cambridge University Press, 1981.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.214

EVOLUTIONARY DEBUNKING ARGUMENTS AND OUR SHARED HATRED OF PAIN

Ben Bramble

WHY DO WE THINK in moral and evaluative terms (i.e., have moral and evaluative beliefs)? According to some philosophers, it is just because such thinking conferred a fitness advantage on our ancestors (i.e., helped them to survive and reproduce) and we have inherited this disposition. It is not because the things that we morally or evaluatively believe are ever true and we are apprehending or otherwise responding to these truths.[1]

In their article, "The Objectivity of Ethics and the Unity of Practical Reason," Katarzyna de Lazari-Radek and Peter Singer argue that while many common moral and evaluative beliefs can indeed be evolutionarily debunked in this way—for instance, the belief that incest between consenting adult siblings is morally wrong—at least one belief cannot be so debunked. This is belief in the following principle:

> *Principle of Universal Benevolence* (UB): each one [of us] is morally bound to regard the good of any other individual as much as his own, except in so far as he judges it to be less, when impartially viewed, or less certainly knowable or attainable by him.[2]

The heart of UB is a claim about what is good *simpliciter*: equal amounts of well-being are worth equally much, no matter whose they are.

Why can belief in UB not be debunked by evolutionary considerations? It is because, say Lazari-Radek and Singer, such a belief would have been no assistance to any of our ancestors in the race to survive and reproduce. On the contrary, it would have been an evolutionary hindrance, for it would have inclined

1 See Richard Joyce, *The Evolution of Morality*; Sharon Street, "A Darwinian Dilemma for Realist Theories of Value."

2 Lazari-Radek and Singer, "The Objectivity of Ethics and the Unity of Practical Reason," 17. This formulation of the principle is due to Henry Sidgwick, *The Methods of Ethics*. Lazari-Radek and Singer's piece is organized around two puzzles raised in Sidgwick's *Methods*. For brevity, I omit discussion here of their interesting claims about Sidgwick's own views on these matters.

one to devote one's time and resources to promoting the survival and reproduction of beings *other* than oneself and one's close kin.

Why, then, do we (or at least many of us) believe UB? According to Lazari-Radek and Singer, a better explanation of this belief is that it is a byproduct of our capacity to reason, which itself was fitness-enhancing. They write:

> [We] might have become reasoning beings because that enabled us to solve a variety of problems that would otherwise have hampered our survival, but once we are capable of reasoning, we may be unable to avoid recognizing and discovering some truths that do not aid our survival.[3]

Having made this case for UB, Lazari-Radek and Singer move swiftly to a dramatic conclusion: utilitarianism is true. By utilitarianism, they mean the view that there is just one ultimate source of normative reasons for action—*well-being, whosever it is*—and so "we ought to maximize well-being generally."[4]

In his piece, "Evolution and Impartiality," Guy Kahane responds to Lazari Radek and Singer. According to Kahane, even if Lazari-Radek and Singer are right that debunking arguments succeed against all beliefs that conflict with belief in UB, but not against belief in UB itself, we can nonetheless evolutionarily debunk *our core beliefs about the nature of well-being*—in particular the beliefs that pleasure is good for us and pain is bad for us. It is obvious, Kahane says, that we have these beliefs about the nature of well-being only due to their usefulness to our ancestors. He writes: "if *any* normative belief can be given such an explanation, it is the universal (or near-universal) conviction that pain is bad [for us]."[5]

Why is this supposed to pose a problem for the argument of Lazari-Radek and Singer? It is a problem, Kahane says, because it would mean that the sort of utilitarianism they had vindicated would be one that is "entirely idle."[6] This is because it would be one on which we can have no idea *what* to maximize.

Kahane adds that some beliefs about well-being might well survive debunking attempts, but only counterintuitive ones like that well-being consists in "ascetic contemplation of deep philosophical truths, or celibate spiritual communion with God, or a kind of Nietzschean perfectionist aestheticism (which might even revel in pain), and so forth."[7] There would have been no evolution-

3 Lazari-Radek and Singer, "The Objectivity of Ethics and the Unity of Practical Reason," 16.

4 Lazari-Radek and Singer, "The Objectivity of Ethics and the Unity of Practical Reason," 18.

5 Kahane, "Evolution and Impartiality," 330–1. For simplicity, in what follows I will focus on the belief that pain is bad for us (though similar points apply to the belief that pleasure is good for us).

6 Kahane, "Evolution and Impartiality," 332.

7 Kahane, "Evolution and Impartiality," 334.

ary benefit to having these beliefs. In this case, the sort of utilitarianism Lazari-Radek and Singer would have vindicated would not be idle, but (worse) it would be one that prescribed perverse actions. Kahane writes:

> If we take the goal of purging all evolutionary influence from our normative views seriously enough, we will end up with a view that is so radically divorced from common sense, and so distant from any familiar ethical theory, that, by comparison, Singer's own utilitarianism will seem almost like old-fashioned common sense.[8]

Kahane concludes that utilitarians like Lazari-Radek and Singer should not invoke evolutionary debunking arguments.

In what follows, I will respond to Kahane on behalf of Lazari-Radek and Singer. I will argue that Kahane's crucial premise (i.e., his claim that a debunking explanation can easily be given of our beliefs that pleasure is good for us and pain is bad for us) is false. On the contrary, these evaluative beliefs are the *hardest* to debunk. The best explanation of them, I will claim, is that we form them as a response to seeing that there is something worth having in pleasure and something worth avoiding in pain. Not only, then, does attention to the origins of our core beliefs about well-being not undermine the strategy of Lazari-Radek and Singer, but it also might vindicate it.

1. THE ARGUMENT FOR KAHANE'S PREMISE

Kahane's key premise is that our core beliefs about well-being are "easy targets for evolutionary debunking."[9] He makes this claim several times. He writes:

> Many of our evaluative beliefs about well-being, including the beliefs that pleasure is good and pain is bad, are some of the most obvious candidates for evolutionary debunking.[10]

And again:

> Our evaluative beliefs about pain and pleasure are perhaps the easiest to explain in evolutionary terms. It will be hard, at best, to find a serious evolutionary theorist who would deny this.[11]

8 Kahane, "Evolution and Impartiality," 330.

9 Kahane, "Evolution and Impartiality," 330.

10 Kahane, "Evolution and Impartiality," 330.

11 Kahane, "Evolution and Impartiality," 334. Kahane does not cite any evolutionary theorists.

And again:

> If evolution has had any role in shaping our evaluative beliefs, it would be hard to deny, to put it mildly, that evolution has played a key role in disposing us to see pain as bad, as something we have strong reason to avoid and minimize.[12]

What is Kahane's argument for this claim? The only argument he offers is one given by Sharon Street. He cites her as follows:

> It is of course no mystery whatsoever, from an evolutionary point of view, why we and the other animals came to take the sensations associated with bodily conditions such as [cuts, burns, bruises, broken bones] to count in favor of what would avoid, lessen, or stop them rather than in favor of what would bring about and intensify them. One need only imagine the reproductive prospects of a creature who relished and sought after the sensations of its bones breaking and its tissues tearing; just think how many descendants such a creature would leave in comparison to those who happened to abhor and avoid such sensations.[13]

It is unclear exactly what the argument here is supposed to be. Of course, a being who *relished and sought after* the sensations of its bones breaking and tissues tearing would not be very fit, evolutionarily speaking. But how does this tend to show that *believing that one's pain is bad for one* would have been fitness-*enhancing*? That wanting and seeking one's pain makes it more likely one will be injured has no obvious tendency to show that believing that one's pain is bad for one makes it less likely that one will be injured.

Let us, then, ignore Street's claim about relishing and seeking after pain, and ask directly: does believing that one's pain is bad for one make it more likely that one will avoid injury?

The answer, it seems to me, is that it would do so *only if one does not already hate or have an aversion to one's own pain*. If one already hates one's own pain, then one already has a considerable motive to avoid such pain. Coming to additionally believe that one's pain is bad for one would seem to add little or nothing to one's tendency to avoid such feelings, and so to one's evolutionary fitness. Indeed, in creatures who already hate their own pain, adding a belief in the badness for them of pain might, at least in certain circumstances, *reduce* their fitness by leading them to focus excessively on immediate pain relief to the exclusion of

12 Kahane, "Evolution and Impartiality," 331.

13 Street, "A Darwinian Dilemma for Realist Theories of Value," 150.

more important things (such as the survival or flourishing of kin or certain long-term interests of their own).

Suppose this is right. A revised debunking argument is nonetheless available to Kahane/Street. According to this argument, while the belief that one's pain is bad for one is not *itself* fitness-enhancing, we have such a belief *only as a result of our hatred of pain*, and we have our hatred of pain only because such hatred was fitness-enhancing in our ancestors. If this is true, then there does appear to be an effective evolutionary debunking argument against the belief that one's own pain is bad for one. Why do we have this belief? Merely as a byproduct of a different state that itself is something we have only for its contribution to our fitness. This, indeed, may be the argument that Kahane/Street had intended to give all along, but failed to put clearly.

2. EVALUATING THE ARGUMENT FOR KAHANE'S PREMISE

What should we make of this revised argument for Kahane's premise? I will not consider the part of it that says that our belief that pain is bad for us is the result of our hatred of pain (though clearly this claim itself needs much fleshing out if the argument is to succeed). I want, instead, to focus on the part of the argument that says that our hatred of pain is due to the usefulness of such hatred in our ancestors. Would such hatred have assisted our ancestors in avoiding injury and so surviving and reproducing?

I think it unlikely that we can explain our hatred of pain in this way. It is certainly true that hating pain *nowadays* (and, more generally, at most times in our recent evolutionary history) is useful in avoiding injury. Young children, especially, who happen not to hate the feeling of pain certainly have a tendency to get injured and die. This is because pain is correlated with injury.

But *why* is pain correlated with injury? Why, in other words, is injury painful in the first place? Kahane and Street seem to assume that it is simply a brute fact that injury is painful. But this might not be so. Perhaps bodily injury was not painful in the beginning at all (i.e., in early beings from whom we are descended), but only *came* to be painful for beings like us as a result of evolution's selecting for beings for whom injury merely happened (i.e., as a result of random mutation) to be painful. Why would evolution have selected for such beings? Perhaps because *all beings start out with a hatred of their own pain*. If all beings start out with a hatred of their own pain, then beings who merely happen to feel pain when they are injured will naturally do much better than their rivals when it comes to avoiding such injury.

On this account, it was hatred of pain (not the painfulness of injury) that was

a given, and was what evolution made use of to get beings to more effectively stay away from things that could injure them. Evolution, in other words, did not make us hate pain by weeding out those who happened to be indifferent to or like it. Instead, it made injury painful by weeding out those for whom it happened not to be painful.

Suppose this is true, that bodily damage is painful for us today only because our distant ancestors had a preexisting hatred of their own feelings of pain. The vital question now is: what explains this preexisting hatred?

The realist has an intriguing answer. She can say that our hatred of pain is in some way *a response to pain's objective badness for us*. Perhaps the badness for one of pain is a normative fact that is so obvious that any creature, even quite basic creatures, can apprehend it (indeed, cannot *avoid* apprehending it). We see that pain is bad for us, and this is why we are all inclined to avoid it.[14] Alternatively, there is some way in which pain's objective badness for us leads us all to hate pain *without* it being necessary that we realize or apprehend its badness for us.[15]

3. A THIRD POSSIBILITY

I have outlined two possible orders of explanation. First, there is that of Kahane/Street, on which the connection between pain and injury is basic, and we have evolved a hatred of pain because such hatred helped our ancestors avoid injury. Second, there is the order of explanation I have proposed, on which it is our hatred of pain that is basic (and in some way a response to pain's objective badness for us), and evolution selected for beings for whom injury merely happened to be painful, thereby leading to injury being painful for all of us today.

But there is a third possibility I have not yet mentioned. This is that neither the connection between pain and injury nor our hatred of pain is basic. Instead, these things evolved *together, as jointly advantageous*.[16] What was fitness-enhancing was hating *something* that was correlated with many different ways in which one might get injured or have one's fitness reduced. Such hatred of a single object (in this case, a feeling produced by each of these threats to one's fitness) was

14 As Irwin Goldstein says in connection with this issue, "creatures of elementary intelligence are still capable of elementary insights" ("Why People Prefer Pleasure to Pain," 357). The suggestion here is not that these animals can necessarily think about value, or have evaluative thoughts in anything like a language like our own, but just that they can be aware on some level, and in some way, that when they are in pain something bad is going on.

15 A further possibility, it must be conceded, is that there is an explanation of our preexisting hatred of pain that appeals neither to its usefulness in evolution nor to its objective badness for us. I will not further consider this possibility here.

16 I am grateful here to an anonymous reviewer for this journal.

more efficient, or "cheaper" in evolutionary terms, than evolving independent desires to avoid each of these different threats. If this third possibility were true, then we would have a successful evolutionary debunking explanation of our belief that pain is bad for us.

What should we make of this hypothesis? I cannot fully address it here, but I believe it is implausible for the following reason. If the sole function of the feeling of pain were to be a feeling we hate, then it seems, at least theoretically, that the feeling of pain could have been *any old feeling*. Any feeling at all might equally well have done the job. But this, intuitively, is not what we find to be the case. The feeling of pain is qualitatively unlike any other feeling, in the following respect: it is one that seems to be in some sense *worth* hating or being averse to. Its aversiveness seems somehow to be built into its very nature.

Now, you might object that this is exactly how we should expect to feel about a feeling that was selected for us to be averse to it. We should, of course, expect to find it aversive. But there is a difference between finding something aversive in the sense merely of wanting it not to occur (the only sense in which, on this third possibility, it should be necessary that we are averse to the feeling selected to be pain), and finding it aversive in the sense in which we find actual pain aversive. Many things we hate, or want not to be the case, are not aversive to us in remotely the way that pain is aversive to us. Why could the feeling of pain not have had a quality more like one of these other things? The peculiar and unique phenomenal quality of pain, I am suggesting, gives us some reason to think that it was not chosen merely as a "thing for us to hate," but that we hate it in some sense because it merits our hatred.

4. CONCLUSION

What is the upshot? Not only does attention to the origins of our core beliefs about well-being not undermine the strategy of Lazari-Radek and Singer, it may well vindicate it. The beliefs that pleasure is good for us and pain is bad for us may have a firmer foundation than any other evaluative beliefs we have—a very firm foundation indeed.

Lazari-Radek and Singer claim that the capacity to reason helped us to evolve and then, *as a side effect*, enabled us to recognize normative truths, like that of utilitarianism. If I am right, however, our ability to recognize normative truths is not only a side effect of the usefulness of our capacity to reason, but also *an essential part of its usefulness*. It was our ability to recognize the goodness of pleasure and the badness of pain in the first place that got us wanting pleasure and

hating pain, the very attitudes that evolution then exploited to direct us toward food and away from bodily injury.

Trinity College Dublin
brambleb@tcd.ie

REFERENCES

Goldstein, Irwin. "Why People Prefer Pleasure to Pain." *Philosophy* 55, no. 203 (July 1980): 349–62.

Joyce, Richard. *The Evolution of Morality*. Cambridge, MA: MIT Press, 2007.

Kahane, Guy. "Evolution and Impartiality." *Ethics* 124, no. 2 (January 2014): 327–41.

Lazari-Radek, Katarzyna de, and Peter Singer. "The Objectivity of Ethics and the Unity of Practical Reason." *Ethics* 123, no. 1 (October 2012): 9–31.

Sidgwick, Henry. *The Methods of Ethics*. 7th ed. London: Macmillan, 1907.

Street, Sharon. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127, no. 1 (January 2006): 109–66.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.215

INTERESTS, WRONGS, AND THE INJURY HYPOTHESIS

Richard Healey

MANY PHILOSOPHERS HOLD that there is an important explanatory and justificatory connection between interests and wrongs.[1] But interest-based accounts of wronging face a significant challenge: we believe many acts constitute wrongs whether or not they set back an individual's interests. Consider a case described by Arthur Ripstein:

> Suppose that, as you are reading this in your office or in the library, I let myself into your home, using burglary keys that do no damage to your locks, and take a nap in your bed. I make sure everything is clean. I bring hypoallergenic and lint-free pajamas and a hairnet. . . . I do not weigh very much, so the wear and tear to your mattress is non-existent. By any ordinary understanding of harm, I do you no harm.[2]

Despite his doing you no harm, we clearly believe Ripstein wrongs you. How, though, can an interest-based theory explain such cases of *harmless wronging*? In *Shaping the Normative Landscape*, David Owens develops a novel solution.[3] Owens argues that, in order to explain harmless wrongs, we must postulate *normative interests*, interests *in* normative phenomena such as *wrongs* and *obligations*. My aim here is twofold. First, I argue against Owens's solution to cases of harmless wronging (sections 1 and 2). Second, and more generally, I show that cases of harmless wronging only pose a problem for interest-based theories if we accept a significant assumption about the relationship between interests and wrongs (section 3).

1 In this article, I am only concerned with the relational form of wrong that involves one agent's wronging another (e.g., *A*'s wronging *B*). I will refer to this as a wrong or wronging interchangeably.

2 Ripstein, "Beyond the Harm Principle," 218.

3 All parenthetical references are to this work.

1. NORMATIVE INTERESTS AND EXPLANATION

Many people recognize the intuitive connection between interests and wrongs. Plausibly, this is because we recognize a connection between making an agent's life go worse and wronging them. This connection would seem to be naturally expressed in what Owens calls the injury hypothesis:[4]

> *Injury Hypothesis* (IH): An act wrongs *X* in virtue of being an action against some interest of *X*'s.[5] (61)

However, cases of harmless wronging pose a problem for IH. According to IH, it is a necessary (although not sufficient) condition on an act's being wrong that it sets back the interests of an individual. But many actions—e.g., Ripstein's harmless trespass—are widely recognized to constitute wrongs that do not set back the interests of any individual.

Owens seeks to solve this problem, and maintain the connection between interests and wrongs, by claiming that we do in fact have an interest in cases of harmless wronging, albeit an interest of a different kind.[6] Specifically, Owens postulates normative interests, interests that take "normative phenomena as [their] object, [interests] *in* which thoughts, feelings, and actions are obligatory, blameworthy, appropriate, or even intelligible" (vi).

In order to make this idea concrete, I focus on cases involving consent (or a lack of it), such as Ripstein's harmless trespass. Here, Owens postulates a *permissive interest*: an interest in being able to "determine by declaration whether something constitutes a wronging" (172). So, for example, Owens holds that the interest that underpins my power of sexual consent is not an interest in having *control* over whether others have sex with me, but is "an interest in its being the case that one is wronged by [sex] *unless one consents to it*" (181).[7] Likewise, I assume, the interest that grounds my power of consent over my property is an interest in its being the case that I will be wronged by the use of my property unless I consent to it. Unlike nonnormative interests, however, normative interests can *only* ground wrongs in the context of a suitable social convention (9–10, 65, 150).

4 Owens claims that the injury hypothesis is "widely accepted" and attributes the view to Joseph Raz, T. M. Scanlon, Judith Jarvis Thomson, and J. S. Mill (61n28).

5 Owens also allows that the adoption of an attitude may set back another's interests. This complication does not bear on my argument, so I set it aside for present purposes.

6 However, Owens does *not* endorse the injury hypothesis himself. See section 3.

7 Importantly, Owens maintains that our normative interests are not reducible to any nonnormative interest—that is, to any more familiar interest in nonnormative phenomena such as control over one's life, pleasure, or knowledge (65).

For example, the harmless trespass will only constitute a wrong within the context of a convention of property rights.

So, Owens claims we can account for cases of harmless wronging if:

(i) We postulate a normative interest (e.g., a permissive interest), and

(ii) A convention instantiates the wrong in question, thereby serving the normative interest (e.g., property conventions instantiate the wrong of trespass, and serve the permissive interest).

However, I do not believe we can offer a meaningful explanation of why ϕ-ing constitutes a wrong—that draws on the intuitive connection between interests and wrongs—by simply postulating an interest in ϕ's being wrong. For an interest of *X*'s to help explain why *Y* may wrong *X* by ϕ-ing, we must recognize the way in which ϕ-ing will (or at least can) make *X*'s life go worse.

Consider two examples. If I claim that *Y* will wrong *X* by setting him on fire I can (partially) explain this by appealing to the fact that setting *X* on fire will cause him to suffer excruciating pain and thereby make his life go much worse. Since we recognize the general claim that making someone's life go worse will often wrong them, and we recognize that *Y*'s setting *X* on fire will make *X*'s life go much worse, we have a clear (if incomplete) explanation as to why *Y* will wrong *X* by setting him on fire. By contrast, imagine that Sally claims that *Y* will wrong *X*—who lives on another continent—by singing a nursery rhyme. We ask Sally why this would be. She says that it is because *X* has an interest in *Y*'s not singing nursery rhymes. But while this seems to fit the argumentative schema (because it appeals to a purported interest of *X*'s that *Y* will set back), it does not offer an illuminating explanation of why *Y* will wrong *X* by singing. That is because we do not recognize how *Y*'s singing could make *X*'s life go any worse. If Sally is to explain or justify her claim, she must do more than simply say that *X* has an "interest" in *Y*'s not singing. She must further show us how *Y*'s singing will make a difference to how *X*'s life goes.

The problem for Owens is that his claim, for example, that a harmless trespass is wrong because we have a permissive interest is in the relevant respects like Sally's claim that *X* has an interest in *Y*'s not singing a nursery rhyme. Put generally: it is not clear why our lives go better *because* certain acts are recognized to constitute wrongings within certain social conventions, independently of the effect these conventions will have on our nonnormative interests. For instance, the idea that an individual's life goes better *just because* it is socially recognized that it is wrong to use her property without her consent is not, I contend, a recognizable human interest. Thus, appealing to the permissive interest to explain cas-

es such as the harmless trespass does not provide the kind of explanation that makes interest-based accounts appealing.

Of course, this is not news to Owens. His strategy is to *postulate* normative interests, and so to *assume*, for the purposes of his discussion, that the social recognition of harmless wrongs makes our lives go better. My point is just that unless we are offered independent grounds for thinking that we *have* normative interests—grounds that are independent of the fact that these interests can explain why harmless wrongs constitutes wrongings—we cannot rely on normative interests to *explain* harmless wrongs. The permissive interest is just, at this stage, a placeholder for the very thing that we seek to explain.[8]

2. Owens's Theory of Wrongs

It might be suggested that, given the difficulty of accounting for harmless wrongs, we should accept this explanatory deficit. However, accepting the postulation of normative interests has further implausible consequences for the structure of Owens's theory of wrongs.

First, on Owens's view, many wrongs are *divorced* from the nonnormative interests we previously took to explain them. This is a consequence of Owens's claim that the *primary* or *central wrong* in any case where there is a conceptually possible harmless instance of that wrong is a harmless wronging. For example, we usually believe that doctors wrong their patients by acting without their consent because we recognize patients' significant *nonnormative* interests in having control over their own lives, bodily integrity, and so forth. But according to Owens, the wrong of operating without consent is instead grounded in a permissive interest. This is because there are conceptually possible harmless instances of this wrong—that is, cases of nonconsensual surgery that are harmless.[9]

If Owens is correct, we must significantly revise our understanding of the nature of these wrongs. But this is counterintuitive for several reasons. First, in the statistically central cases of these wrongs, victims have weighty *nonnormative* interests that are set back, and it seems plausible that it is the significance of these nonnormative interests that primarily motivates our concern with these cases.[10]

8 See further Bennett, "A Review of David Owens' *Shaping the Normative Landscape*," and Markovits, "Authority, Recognition, and the Grounds of Promise."

9 This might be disputed, but I am assuming it here for the sake of argument.

10 Owens maintains, against this, that cases of harmless wronging show that, "For analytical purposes, the central cases … are the statistically peripheral ones in which no harm aggravates the primary wrong" (177). This claim gives rise to interesting methodological questions that I cannot address here.

Furthermore, divorcing all consent-related wrongs from nonnormative interests has the implication that the wrong of acting without consent always depends upon the existence of an appropriate *social convention*, because, as we saw above, normative interests only ground wrongs in the presence of social conventions. Thus, as Owens notes, his theory has the striking consequence that "breach of promise, disloyalty in friendship, and even rape (in abstraction from the harm it causes) constitute wrongs only against a background of social convention" (9). Yet, while perhaps not implausible in all cases, surely the central egregious wrongs of rape, or of nonconsensual human experimentation, do not depend on the existence of social conventions.[11]

Owens may respond by pointing out that he is clear that there are usually *secondary* and *convention-independent* wrongs grounded in nonnormative interests. While not related to our power of consent, Owens claims that these secondary wrongs are affected by the normative significance of the choices we make (ch. 7). Yet this response leads to the second counterintuitive structural implication of Owens's view: in the statistically central cases of many wrongs, there is a *doubling* of wrongs. That is because there will generally be a wrong grounded in a normative interest *as well as* a wrong grounded in a nonnormative interest. For example, a surgeon may perpetrate both a *harmless* wronging, grounded in the patient's permissive interests, as well as a harmful wronging, grounded in the patient's nonnormative interests. Yet, claiming there are two wrongs in such cases is implausible. Imagine that after an operation a patient complains that the surgeon has wronged him because, despite giving consent to undergo the operation, he did not *choose* the surgery.[12] Surely the patient is mistaken: if he gave free and informed consent to the operation then the surgeon has not wronged him because he did not also choose the surgery.[13] Yet for Owens, if we assume the truth of the patient's claim, he is correct (181).

11 Of course, conventions of consent may well shape what morally permissible relations consist in. Conventions of sexual consent may, for example, serve to make what constitutes morally permissible sexual relations more determinate in different contexts. The point is just that, as participants in the ongoing debate about how to characterize the central wrong of rape would seem to agree, the basic wrong of rape is not essentially conventional.

12 *Choosing* the surgery, on Owens's rendering, requires him to intend for the surgery to go ahead (173).

13 To be clear, I am not claiming that there is no way that the patient could have been wronged by consensual surgery. They could have been so wronged if, for example, the surgeon violated a fiduciary duty. What I am claiming is that they could not have been wronged by surgery that they *consented* to even if they did not *choose* it.

3. THE INJURY HYPOTHESIS

Owens may respond by suggesting that if we are committed to an interest-based account of wronging, we must postulate normative interests in order to account for harmless wrongs, even if this means accepting some seemingly counterintuitive results. Alternatively, we may be tempted to give up on the intuitive connection between interests and wrongs altogether. Indeed, Ripstein introduces the example of harmless trespass in order to motivate a move away from Mill's *harm principle* to an alternative *sovereignty principle*.[14]

However, we are only forced to choose between postulating normative interests—and accepting the difficulties with Owens's theory that I have highlighted—or rejecting the intuitive connection between interests and wrongs altogether, *if* we have reason to endorse IH. That is because it only follows that interest-based theories are unable to account for harmless wrongings if we accept as a necessary condition on an act's being wrong that it must set back an interest of the victim. Do we, then, have reason to endorse IH? Since I cannot consider the plausibility of IH in detail here, I will defend a relatively weak claim: those who give nonnormative interests a central role in their theory of wronging are not *required* to endorse IH.[15] That is to say, there is conceptual space within which to develop accounts of the relationship between interests and wrongs that do not incorporate IH.[16] Indeed, Owens himself rejects IH (64). But while Owens sometimes gives the impression that his rejection of IH is connected to

14 Ripstein, "Beyond the Harm Principle."

15 As an interpretive matter, I do not think it is clear that all the authors Owens cites as holding the view (viz., Raz, Scanlon, Thomson, and Mill) really do hold it. I cannot pursue this interpretative question here.

16 Defenders of interest-based theories might also argue that, in the cases of purportedly "harmless" wronging under consideration, we in fact have noninstrumental interests in having control over what happens (to us or to our property), and that these interests will be set by the act in question. Importantly, on this view, our control interests will be set back by the wrongdoer's action independently of whether this will have any further effect on how our life goes. If this claim can be defended, this would allow us to accept IH while accounting for cases of ostensibly harmless wronging. However, while I am sympathetic to the idea that we have such interests in many cases (e.g., with regard to our sexual relationships), I suspect that our conviction that harmless wrongs are genuine wrongs does not depend upon our belief that we can find an interest that will be set back in every proposed case of harmless wronging. Rather, I think that our beliefs about these cases flow from a more deep-rooted (albeit inchoate) view about the nature of wrongs, according to which the correct account of the relationship between interests and wrongs does not incorporate IH. Thus, one of my main aims in this article is simply to challenge IH, leaving a full consideration of other argumentative strategies for another occasion. Thanks go to an anonymous reviewer for urging me to consider this possibility.

the postulation of normative interests, there is no reason to think that we must postulate normative interests before we are entitled to reject IH.

To see this, note first that IH implicitly depends upon a form of consequentialism about wronging: according to IH, if *Y* is to wrong *X* by acting it must be the case that *Y*'s action is an action against some interest of *X*'s. I take it that this means that the action must *cause the setting back* of *X*'s interest; if *Y*'s act does not cause the setting back of *X*'s interest then it fails to bring about a *consequence* that is a necessary condition upon an act's being wrong. If this is correct, however, IH depends on controversial consequentialist assumptions about the relationship between values (in this case wellbeing) and wrongs.[17] Rather than postulating normative interests, or rejecting interest-based theories outright, we may prefer to reject the implicit consequentialism that forces this choice upon us.[18]

Unsurprisingly then, in order to show that a conception of wrongs need not incorporate IH, we need look no further than a *nonconsequentialist* theory of wrongs. For illustrative purposes, consider T. M. Scanlon's popular account of contractualism. According to Scanlon, an act is wrong if its performance would be disallowed by principles that no one could reasonably reject.[19] Whether Scanlon's principle implicitly endorses IH is an open question. But there is good reason to think that it need not. Indeed, Rahul Kumar has argued that part of the *appeal* of Scanlonian contractualism is that nothing like IH holds.[20] On Kumar's reconstruction, principles that could not be reasonably rejected serve to "fix what the general terms are for relating to one another … by establishing certain legitimate expectations concerning considerations and conduct between persons."[21] We *wrong* one another, on this view, if we fail to abide by these *legitimate expectations*.

What is crucial for our purposes is the point at which interests can enter the foregoing story—namely, as grounds for the reasonable rejection of principles. For instance, I may reasonably reject a principle that licenses others to use my bed when they wish because of my interests in autonomy, privacy, etc. As such, I may form legitimate expectations that others refrain from breaking into my

17 See Kumar, "Defending the Moral Moderate," 278–79, and the references therein, for discussion of a closely related point.

18 I am not making the strong claim that consequentialist theories *must* accept IH. I merely want to make clear that IH would seem to presuppose a *form* of consequentialism about wronging. It does seem, however, that nonconsequentialists will have more resources to motivate views that do not endorse IH.

19 Scanlon, *What We Owe to Each Other*, 153.

20 See Kumar, "Defending the Moral Moderate" and "Who Can Be Wronged?"

21 Kumar, "Who Can Be Wronged?" 106.

house and using my bed. When others violate these legitimate expectations, they wrong me by doing so *whether or not* they set back any of my interests on this occasion. What matters is that, given my interests, I can have certain legitimate expectations of others, and I will be wronged when they violate these expectations. On this account, then, the connection between interests and wrongs is *indirect*: whether *Y* wrongs *X* depends upon whether *Y* violates a principle that *X* could reasonably reject; whether *X* can reasonably reject a principle will depend (in many cases) upon *X*'s interests.

In summary: I have argued that Owens's solution to cases of harmless wronging fails to be explanatory, and has implausible consequences for his theory of wrongs. Furthermore, I have shown that the argumentative significance of these cases requires that we accept a substantive and questionable assumption about the relationship between interests and wrongs.[22]

McGill University and Centre de Recherche en Éthique
richard.healey@mail.mcgill.ca

REFERENCES

Bennett, Christopher. "A Review of David Owens' *Shaping the Normative Landscape*." *Jurisprudence* 6, no. 2 (2015): 364–70.

Kumar, Rahul. "Defending the Moral Moderate: Contractualism and Common Sense." *Philosophy and Public Affairs* 28, no. 4 (Autumn 1999): 275–309.

———. "Who Can Be Wronged?" *Philosophy and Public Affairs* 31, no. 2 (Spring 2003): 99–118.

Markovits, Daniel. "Authority, Recognition, and the Grounds of Promise." *Jurisprudence* 6, no. 2 (2015): 349–56.

Owens, David. *Shaping the Normative Landscape*. Oxford: Oxford University Press, 2012.

Ripstein, Arthur. "Beyond the Harm Principle." *Philosophy and Public Affairs* 34, no. 3 (Summer 2006): 215–45.

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

22 Thanks go to Jason D'Cruz, Jens Gillessen, Angie Pepper, Sarah Stroud, and an anonymous referee for *JESP* for helpful comments on an earlier version of this article.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.216

OUR INTUITIONS ABOUT THE EXPERIENCE MACHINE

Rach Cosker-Rowland

Felipe de Brigard, Adam Kolber, Wayne Sumner, Dan Weijers, and Katarzyna de Lazari-Radek and Peter Singer have argued that our intuitions about Nozick's experience machine are untrustworthy because they are distorted by biases and irrelevant factors. De Brigard and Weijers recently conducted empirical studies regarding people's intuitions about versions of the experience machine to test which of our intuitions are not distorted by such biases and irrelevant factors. They claim their results show that our intuitions about the experience machine do not undermine hedonism (section I). I argue, on the basis of further empirical studies, that De Brigard and Weijers fail to establish that our intuitions about the experience machine do not undermine hedonism (section II).

Hedonism is the view that the only thing that is noninstrumentally good *for* us or that noninstrumentally improves our well-being is pleasure. Nozick famously asked us to consider the following thought experiment:

> Suppose there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences?[1]

Nozick presumed that upon considering the experience machine the vast majority of people would judge that they should choose to remain in reality rather than plug into the machine, and that this gives us strong reason to reject hedonism. And our intuitions about Nozick's experience machine are widely held to give us strong reason to reject hedonism.

1 Nozick, *Anarchy, State, and Utopia*, 42.

1

De Brigard, Kolber, Sumner, Weijers, and Lazari-Radek and Singer argue that our intuitions about Nozick's experience machine are prone to some of the following misleading factors and biases:

(a) Imaginative resistance to the possibility of such an experience machine, such as worries that the machine might not work as well as we think it does.
(b) The fact that we find it hard to give up responsibility for our loved ones in the way that we must if we plug into the experience machine.
(c) Overactive imaginations overreacting to the scenario and being overly horrified by Nozick's characterization.
(d) Status quo bias—that is, an inappropriate or irrational preference for an option because it preserves the status quo.[2]

According to these philosophers, the fact that our intuitions about the experience machine are prone to (a)–(d) directly or indirectly establishes that these intuitions do not give us strong reason to reject hedonism.[3]

This argument is made most strongly and plausibly by De Brigard and Weijers, who each develop versions of the experience machine our intuitions about which are *supposedly* not prone to (a)–(d). Call De Brigard's and Wejiers's versions of the experience machine *Undistorted Experience Machines*. De Brigard and Weijers conducted studies of people's intuitions about *Undistorted Experience Machines* and Nozick's experience machine. They found that, although a majority judge that we should choose reality rather than Nozick's experience machine, it is not the case that a majority judge that we should choose reality rather than an *Undistorted Experience Machine*. De Brigard and Weijers conclude that this shows that our intuitions about the experience machine do not give us strong reason to reject hedonism.

In De Brigard's *Undistorted Experience Machine* cases, we are asked to imagine being told by a reliable source that we have been living life inside an experience machine and are given the chance to unplug to lead a different life in reality. De Brigard's cases are intended to prevent status quo bias from influencing our

2 See Bostrom and Ord, "The Reversal Test: Eliminating Status Quo Bias in Applied Ethics."

3 See De Brigard, "If You Like It, Does It Matter If It's Real?"; Kolber, "Mental Statism and the Experience Machine," 13–16; Sumner, *Welfare, Happiness and Ethics*, 21; Weijers, "Intuitive Biases in Judgments about Thought Experiments" and "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" 517–19; and Lazari-Radek and Singer, *The Point of View of the Universe*, 254–61.

anti-experience-machine intuitions.[4] But they succeed in this endeavor at the cost of rendering pro-experience-machine intuitions prone to status quo bias; for in De Brigard's cases, the status quo is the experience machine.[5]

Weijers's *Undistorted Experience Machine* avoids this problem. In Weijers's *Undistorted Experience Machine*, we are asked to consider the following case:

> A stranger, named Boris, has just found out that he has been regularly switched between a real life and a life of machine-generated experiences (without ever being aware of the switches); 50% of his life has been spent in an Experience Machine and 50% in reality. Nearly all of Boris' most enjoyable experiences occurred while he was in an Experience Machine and nearly all of his least enjoyable experiences occurred while he was in reality. Boris now has to decide between living the rest of his life in an Experience Machine or in reality (no more switching).
>
> You have had a go in an Experience Machine before and know that they provide an unpredictable rollercoaster ride of remarkable experiences. When in the machine, it still felt like you made autonomous decisions and occasionally faced tough situations, such as striving for your goals and feeling grief, although you didn't really do these things. Your experiences were also vastly more enjoyable and varied in the machine. You also recall that, while you were in the Experience Machine, you had no idea that you had gotten into a machine or that your experiences were generated by a machine.
>
> Boris' life will be the same length in an Experience Machine as it would in reality. No matter which option Boris chooses, you can be sure of two things. First, Boris' life will be very different from your current life. And second, Boris will have no memory of this choice and he will think that he is in reality.
>
> (1) Ignoring how Boris' family, friends, any other dependents, and society in general might be affected, and assuming that Experience Machines always work perfectly, what is the best thing for Boris to do for himself in this situation? Tick only one of these options:
>
> ☐ Boris should choose the Experience Machine life.
> ☐ Boris should choose the real life.
>
> (2) Briefly explain your choice.[6]

4 De Brigard, "If You Like It, Does It Matter If It's Real?" 47–50.

5 Weijers, "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" 519.

6 Weijers, "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" 525–26.

Weijers gave eighty-two subjects his *Undistorted Experience Machine*. Of those eighty-two, 54.5 percent said that Boris should choose the experience machine life.[7] Weijers also asked seventy-nine subjects about Nozick's original version of the experience machine and found that, in comparison, only 16 percent would plug into Nozick's machine.[8] On the basis of this survey data, Weijers concludes that our intuitions about the experience machine do not give us strong reason to reject hedonism.[9] For once we eliminate factors that distort our intuitions about the experience machine, a majority of people do not have the intuition that a life in reality is better *for* us than a life in the experience machine.[10]

II

In Weijers's *Undistorted Experience Machine*, all other things are not held equal between the experience machine life and the real life: life in the experience machine is far more pleasurable and exciting than real life. Given that all things are not held equal in Weijers's case, it is consistent with Weijers's results that all 54 percent—or at least a significant proportion of the 54 percent—who responded to Weijers's vignette by saying that Boris should choose the experience machine life believe, or are inclined to believe:

> *Pleasure-Weighted Non-Hedonism*: Pleasure is not all that is noninstrumentally good *for* us; other things, such as living a life in reality, are noninstrumentally good *for* us too. But a life of great pleasure that is not in reality is better *for* us than a life of far less pleasure in reality because the noninstrumental goodness *for* us of pleasure is greater than the noninstrumental goodness *for* us of any other good.

As I explain, *Pleasure-Weighted Non-Hedonism* is a plausible view that many people hold.

To see the plausibility of *Pleasure-Weighted Non-Hedonism*, suppose that you lead an utterly miserable life of pain and suffering, alone, achieving nothing. You know that you cannot escape this life of torment and pain while remaining in reality. But you have the ability to plug into an experience machine that will give you the immensely enjoyable experience of living a very pleasurable life. We can certainly think that living a life in reality is noninstrumentally good *for* us, but that a life in reality in this case is so bad that it would be preferable and better *for*

7 Weijers, "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" 527.

8 Weijers, "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" 520.

9 Weijers, "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" 529.

10 Weijers, "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" 528–29.

us to plug into the experience machine. And some of De Brigard's experimental results favor the view that many hold *Pleasure-Weighted Non-Hedonism*. For 87 percent of De Brigard's participants preferred staying in the experience machine to a life full of very little pleasure (in a maximum security prison) but only 50 percent of people preferred a life in the experience machine to a very pleasurable life (as a millionaire in Monaco). The combination of De Brigard's results favors the view that many hold *Pleasure-Weighted Non-Hedonism* because the view that many hold that the goodness *for* us of pleasure is greater than the goodness *for* us of contact with reality. This can explain why so many more people prefer the experience machine to the unpleasant prison life than preferred the experience machine life to the pleasurable millionaire life.[11]

So it seems that if we want to truly test whether people's intuitions about the experience machine count against hedonism, we must investigate their intuitions about an *Undistorted Experience Machine* case in which all things are held equal between the two lives we are able, or another is able, to choose. (Many other philosophers, including Russ Shafer-Landau in his widely used introductory text, and prominent hedonists such as Roger Crisp, also focus on versions of the experience machine thought experiment in which all things are held equal between the experience machine life and the life in reality.)[12] I developed such an *Undistorted Experience Machine* and ran an experimental survey using it.

I recruited eighty-one subjects from Amazon's Mechanical Turk platform. I gave the subjects a variation of Weijers's vignette in which a life in the experience machine was stipulated as being predictably only as enjoyable as a life in reality. I gave subjects the following vignette and questions (italics denote changes from Weijers's vignette; these italics were not in the version subjects were given):

> A stranger, named Boris, has just found out that he has been regularly switched between a real life and a life of machine-generated experiences (without ever being aware of the switches); 50% of his life has been spent in an Experience Machine and 50% in reality. *50% of Boris' most enjoyable experiences occurred while he was in an Experience Machine and 50% of his most enjoyable experiences occurred while he was in reality*. Boris now has to decide between living the rest of his life in an Experience Machine or in reality (no more switching*). Living in the Experience Machine rather than in reality will not make Boris happier and Boris knows this before making his*

11 De Brigard, "If You Like It, Does It Matter if It's Real?" 47–49.

12 See Crisp, *Reasons and the Good*, 118, and Shafer-Landau, *The Fundamentals of Ethics*, 34–35. For a defense of using such an otherwise identical lives version of the Experience Machine, see Lin, "How to Use the Experience Machine."

choice; if Boris chooses a life outside of an Experience Machine his life will be as happy as his life would have been had he chosen a life in an Experience Machine. If Boris does not choose one of these two new forms of life he will not continue living. Boris must make an active choice between these two new lives. And he will not remember having made his choice once he begins his new life.

You have had a go in an Experience Machine before and know that they *are extraordinarily safe and can provide remarkable experiences*. When in the machine, it still felt like you made autonomous decisions and occasionally faced tough situations, such as striving for your goals and feeling grief, although you didn't really do these things. *You felt that your life, plans, relationships, and achievements were real.* You also recall that, while you were in the Experience Machine, you had no idea that you had gotten into a machine or that your experiences were generated by a machine.

Boris' life will be the same length in an Experience Machine as it would in reality. No matter which option Boris chooses, you can be sure of *four* things. First, Boris' life will be very different from your current life. Second, Boris will have no memory of this choice and he will think that he is in reality. *Third, the experience machine will not malfunction or break when Boris is in it. Fourth, Boris' life will be as long and happy as it would have been had he made the alternative choice.*

(1) Ignoring how Boris' family, friends, any other dependents, and society in general might be affected, and assuming that Experience Machines always work perfectly, what is the best thing for Boris to do for himself in this situation? Tick only one of these options:

☐ Boris should choose the Experience Machine life.
☐ Boris should choose the real life.

(2) Briefly explain your choice.

Of the eighty-one subjects, seventy-three—more than 90 percent—said that Boris should choose the real life; only eight subjects—less than 10 percent—said that Boris should choose the experience machine life. (Furthermore, seven of the eight subjects who said that Boris should choose the experience machine life explained their choice in a way that showed that they had either accidentally ticked the option that they did not mean to tick or believed that life in the experience machine would be safer or more pleasurable.)[13]

13 An anonymous referee points out that a convinced hedonist might want to choose a third option that was not offered—namely that it does not matter which life Boris chooses. However, participants were asked to explain their choice. Of the seventy-three subjects who said

So, De Brigard and Weijers are mistaken that experimental data regarding our intuitions about the experience machine establish the failure of the argument against hedonism from our intuitions about the experience machine, since, when all other things were held equal, more than 90 percent of subjects judged that we should, for the sake of what is good *for* us, choose a life in reality over a life in the experience machine.[14] Furthermore, the argument that our intuitions about the experience machine do not give us strong reason to reject hedonism because our intuitions about the experience machine are subject to distorting biases seems to fail too. For when all these distorting and biasing factors were mitigated and all other things were held equal, the vast majority of people judged that we should, for the sake of what is good *for* us, choose a life in reality over a life in the experience machine.[15]

Australian Catholic University
r.cosker-rowland@leeds.ac.uk

REFERENCES

Bostrom, Nick, and Toby Ord. "The Reversal Test: Eliminating Status Quo Bias in Applied Ethics." *Ethics* 116, no. 4 (July 2006): 656–79.

Crisp, Roger. *Reasons and the Good*. Oxford: Oxford University Press, 2006.

---

that Boris should choose the real life, only one said that they thought it made no difference which life Boris chose, such that they would have ticked a third box saying that Boris should flip a coin, if such an option had been provided. The fact that (i) only one of the seventy-three subjects who ticked the "Boris should choose the real life" option said that they would have chosen a "makes no difference" option combined with the fact that (ii) most subjects elucidated a view about how reality matters in itself in their explanation of their choice and the fact that (iii) we would expect subjects who thought it made no difference which life Boris picks to tick the reality or experience machine options with approximately equal frequency, provides good evidence that very few subjects would have chosen a "makes no difference" option if it had been presented.

14 Weijers, "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" 528–29, and De Brigard, "If You Like It, Does It Matter if It's Real?" 44.

15 I would like to thank the University of Oxford's John Fell Fund for generously funding this research. I would also like to thank Lucius Caviola, Dan Cohen, Roger Crisp, Guy Kahane, David Killoren, Tyler Paytas, and Dan Weijers for their advice and/or feedback regarding this article, the ideas within, and/or how to conduct the study for this article. And I would like to thank Sam Billington, whose undergraduate essay at Somerville got me thinking about the issues of this article.

De Brigard, Felipe. "If You Like It, Does It Matter If It's Real?" *Philosophical Psychology* 23, no. 1 (2010): 43–57.

Kolber, Adam J. "Mental Statism and the Experience Machine." *Bard Journal of Social Sciences* 3 (1994): 10–17.

Lazari-Radek, Katarzyna de, and Peter Singer. *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford: Oxford University Press, 2014.

Lin, Eden. "How to Use the Experience Machine." *Utilitas* 28, no. 3 (September 2016): 314–32.

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

Shafer-Landau, Russ. *The Fundamentals of Ethics*. 2nd ed. Oxford: Oxford University Press, 2012.

Sumner, L. W. *Welfare, Happiness and Ethics*. Oxford: Oxford University Press, 1996.

Weijers, Dan. "Intuitive Biases in Judgments about Thought Experiments: The Experience Machine Revisited." *Philosophical Writings* 41, no. 1 (2011): 17–31.

———. "Nozick's Experience Machine Is Dead, Long Live the Experience Machine!" *Philosophical Psychology* 27, no. 4 (2014): 513–35.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.217

THE VALUE OF CARING

A REPLY TO MAGUIRE

Jörg Löschke

One standard objection against classical act consequentialism is that it cannot justify partiality to our loved ones. Classical act consequentialism claims that agents are always required to bring about the impersonally best outcome, or, to put it in more technical terms, that an agent is required to do what she has most reason to do, and that the strength of a reason for action is a function of the value that the action would bring about. The more value an action would realize, the stronger the reason to perform that action, and the action that maximizes value is therefore always supported by the strongest reasons. And classical act consequentialism understands value as agent-neutral value: the goodness of an outcome does not depend on the point of view or the identity of the agent. Agents are therefore not permitted to attach special significance to their own personal relationships when they deliberate about what to do. Of course, relationships might be impersonally valuable, and a world with loving relationships might be better than a world without such relationships. If that is true (and it is plausible to assume that it is), then agents must take the value of relationships into account when they deliberate about what to do. But on an agent-neutral understanding of value, this just means that they are required to regard their own relationships as just as valuable as the relationships of other people. The fact that some relationship is mine bears no special significance.[1]

To illustrate this point, consider the following example that has been put forward by Barry Maguire:

> Angela and Becca both love their mothers. Angela's mother is a beneficent soul, always doing her best to improve the lives of those around her. Becca's mother is rather self-centered, and never really thinks about

1 For different formulations of the criticism that consequentialism cannot accommodate partiality to loved ones, see, among others, McNaughton and Rawling, "Honoring and Promoting Value"; Jollimore, *Friendship and Agent-Relative Morality*; Card, "Consequentialism, Teleology, and the New Friendship Critique"; and Wallace, "Reasons, Values and Agent-Relativity."

> or does much for others. Alas, the two mothers are stuck in a burning building. There is only one oxygen mask; only one daughter can save her mother. Becca is holding the mask. . . . Becca is in a position either to save her mother or enable Angela to save Angela's slightly more beneficent mother.[2]

If agents are always required to bring about the agent-neutrally best outcome, then Becca ought to enable Angela to save Angela's mother. After all, given Angela's mother's character, her surviving would be better from an impersonal point of view. And Angela has the same kind of valuable relationship with Angela's mother as Becca has with Becca's mother, so there is no reason why classical consequentialism should permit Becca to favor her own mother, due to their valuable relationship. However, most people probably share the intuition that Becca ought to save her own mother rather than enable Angela to save Angela's mother. The question is whether consequentialism can accommodate this intuition without sacrificing its core commitments: the agent-neutral conception of value, the teleological conception of reasons (according to which practical reasons are reasons to bring about valuable states of affairs), and the maximizing view (according to which agents are required to bring about the impersonally *best* outcome).

Authors with consequentialist leanings have responded to this challenge from partiality in different ways.[3] However, all of these attempts have met criticism.[4] An interesting new attempt to justify partiality within a consequentialist framework has been recently put forward by Barry Maguire.

Maguire's crucial move is to claim that the strength of an agent's reason to bring about an outcome can be modified by the (agent-neutral) value of the agent's caring about that outcome.[5] It is good that Becca cares about the out-

2 Maguire, "Love in the Time of Consequentialism," 4.

3 Some authors abandon the agent-neutral conception of value and adopt an agent-relative conception of value instead. See, for example, Smith, "Neutral and Relative Value After Moore" and "Two Kinds of Consequentialism," or Portmore, *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Others reject the view that agents are always required to bring about the best possible outcome and adopt a satisficing view, according to which agents are required to bring about outcomes that are sufficiently good. See, for example, Slote, "Satisficing Consequentialism" and *Common-Sense Morality and Consequentialism*, or Vallentyne, "Against Maximizing Act-Consequentialism."

4 For objections against the strategy to rely on agent-relative value, see Schroeder, "Teleology, Agent-Relative Value, and 'Good,'" or Cullity, "Neutral and Relative Value." For problems of satisficing consequentialism, see Mulgan, "Slote's Satisficing Consequentialism," or Bradley, "Against Satisficing Consequentialism."

5 Maguire, "Love in the Time of Consequentialism," 10.

come in which her mother survives, and this means that the state of affairs [Becca's mother survives, and Becca cares about it] is agent-neutrally better than the state of affairs [Angela's mother survives] or the state of affairs [Angela's mother survives, and Becca cares about it].[6] Becca's reason to save her own mother is therefore stronger than her reason to enable Angela to save Angela's mother, even if Angela's mother surviving is the agent-neutrally better outcome, compared to Becca's mother surviving. And since Becca's reason to save her own mother is stronger than her reason to enable Angela to save Angela's mother, Becca ought to save her own mother.

This proposal is interesting, but it fails to justify partiality within a consequentialist framework because it does not explain the special normative significance of an agent's caring attitude *for that agent*. Or so I will argue.

1

To see why Maguire's proposal does not succeed, we can imagine that Becca is a convinced consequentialist and tries to figure out what she ought to do when she faces the choice of either saving her own mother or enabling Angela to save Angela's mother instead. Becca knows that it would be impersonally better if Angela's mother survived, but she also knows that it is good for participants in personal relationships to care about each other—she knows, in other words, that there is value in her caring about the outcome in which her own mother survives. So Becca might deliberate as follows:

> I can either save my own mother, or enable Angela to save Angela's mother. The survival of my own mother and the survival of Angela's mother are both valuable outcomes, so I have reason to bring about either one of them. But Angela's mother is a beneficent soul, and my own mother is rather self-centered. That means that, from an impersonal point of view, it would be better if Angela's mother survives, and I have more reason to enable Angela to save Angela's mother. However, I care about the outcome in which my own mother survives—much more than I care about the

6 Maguire does not put his view in terms of such states of affairs. However, it is natural to interpret him in this way. According to Maguire, the bearers of value are states of affairs. The value of caring about an outcome must therefore also be understood as a valuable state of affairs. Furthermore, Maguire understands the weight of a reason as "a monotonically increasing function of the value of its object state, perhaps along with other variables" ("Love in the Time of Consequentialism," 4). It is therefore plausible to assume that Becca's reason to save her own mother increases in strength due to this additional value in the more fine-grained description of the relevant state of affairs [Becca's mother survives, and Becca cares about it], compared to [Becca's mother survives].

> survival of Angela's mother. And clearly, it is good if children care about their parents in the way that I care about my mother. And my caring about my mother is good not only from my point of view, but agent-neutrally good. So I have to take this value into account when I decide what to do. I do not have to compare [Angela's mother survives] and [Becca's mother survives], but [Angela's mother survives] and [Becca's mother survives, and Becca cares about it]. And if I compare *these* states of affairs, the additional agent-neutral value of my caring about the outcome in which my mother survives makes a difference; it makes [Becca's mother survives, and Becca cares about it] agent-neutrally better than [Angela's mother survives]. It therefore increases the strength of my reason to save my own mother. My reason to save my own mother is stronger than my reason to enable Angela to save Angela's mother—so I ought to save my own mother, rather than enable Angela to save Angela's mother.

At this point, the problem for Maguire's proposal emerges. Maguire accepts that all value is agent-neutral value.[7] He therefore accepts that Becca's caring attitude is agent-neutrally good, and that [Becca's mother survives, and Becca cares about it] is also agent-neutrally good. Agent-neutrally good states of affairs give *all* agents reason to promote them—this is consequentialism's interpretation of impartiality, and it is sometimes taken to be an attractive feature of the theory.[8] This means that not only Becca but also Angela has reason to bring about [Becca's mother survives, and Becca cares about it]. And importantly, it also means that Becca has reason to bring about [Angela's mother survives, and Angela cares about it], since that state of affairs is also agent-neutrally good. We therefore have to compare not the agent-neutral value of [Becca's mother survives, and Becca cares about it] and [Angela's mother survives] but the agent-neutral value of [Becca's mother survives, and Becca cares about it] and [Angela's mother survives, and Angela cares about it].

The problem is that the agent-neutral value of [Angela's mother survives, and Angela cares about it] should be greater than the agent-neutral value of [Becca's mother survives, and Becca cares about it]. Angela's mother surviving is agent-neutrally better than Becca's mother surviving, and there is no reason to think that Becca's caring attitude toward her mother is agent-neutrally better than Angela's caring attitude toward Angela's mother. So if we add the agent-neutral value of the respective caring about the outcome to the agent-neutral value of the respective outcome, we still get the result that [Angela's mother survives,

7 Maguire, "Love in the Time of Consequentialism," 3.

8 See, for example, McElwee, "Impartial Reasons, Moral Demands," 458.

and Angela cares about it] has more agent-neutral value than [Becca's mother survives, and Becca cares about it]. Becca has therefore more reason to bring about the first outcome than she has reason to bring about the second outcome, and that means that she ought to enable Angela to save Angela's mother rather than save her own mother. The value of her caring attitude cannot justify partiality within an agent-neutral consequentialist framework.

Note that this argument does not rely on any specific account of how the value of an agent's caring about an outcome can modify the strength of her reason to bring that outcome about. It does not matter whether we think that the value of an agent's caring about an outcome must be added to the value of that outcome, or if we think that the modification works in some other way.[9] Once we assume that it is the agent-neutral value of my caring about an outcome that modifies the strength of my reason to bring about that outcome, we need an argument for why the agent-neutral value of some other agent's caring about a different outcome does not change my normative situation as well. And as far as I can see, Maguire does not provide such an argument.

One might argue that there is an important difference between *my* caring about a specific outcome and *your* caring about a different outcome: the first caring attitude is mine, and the second is not. One might think that this difference is relevant, and that the value of Becca's caring about a specific outcome modifies the strength of Becca's reason to bring about that outcome, while the value of Angela's caring about a specific outcome modifies the strength of Angela's reason to bring about that outcome, and so on. In this view, the value of Becca's caring attitude is normatively relevant only for Becca and not for Angela, and the value of Angela's caring attitude is normatively relevant only for Angela but not for Becca. It would then follow that Becca ought to save her own mother rather than enable Angela to save Angela's mother, because the weight of Becca's reason to save her own mother is modified in a way that the strength of her reason to enable Angela to save Angela's mother is not.[10]

But what should explain this normative specialness of *my* caring attitude for

9 At one point ("Love in the Time of Consequentialism," 7), Maguire suggests that the strength of a reason is modified by adding the value of the agent's caring about an outcome to the value of that outcome. Later ("Love in the Time of Consequentialism," 14), Maguire seems to suggest that the modification works in some other way, without going into detail of how the modification works.

10 This is what Maguire seems to have in mind when he writes, "The trick is that the partial orientations we are evaluating are the agent's own. We are not evaluating, for instance, the agent's partial orientation towards some states of affairs and someone else's partial orientation towards some state of affairs. It is plausibly of value to be partially oriented towards some outcomes rather than others. This is what makes it possible—even though our expla-

me? In an earlier presentation of the account, Maguire explains the normative specialness of an agent's caring attitude for that agent by arguing that

> attitudes always consist in a relation between an agent and a state of affairs. The agent herself cannot but stand in the first place of this relation. States of affairs consisting in some agent standing in an attitudinal relation to some other state of affairs are regular old states of affairs like any other, and admit of some neutral value or disvalue just like any other. However they are in one important and non-axiological sense agent-relative, simply because the agent cannot remove herself from the first relatum. It is metaphysically impossible for me to have your attitudes, or you mine.[11]

Maguire seems to explain the normative specialness of *my* caring attitude by its descriptive specialness: my attitude is agent-relative in a descriptive way and, from this descriptive specialness, Maguire infers that it is also agent-relative in a normative way. In other words, he seems to assume that, from the fact that *my* caring attitude cannot be anyone else's attitude but *mine*, it follows that *my* caring attitude cannot be normatively relevant for anyone but *me*.

Of course, it is true that an attitude is agent-relative in the sense that its existence depends on the agent who has that attitude. Becca can only have Becca's attitudes, and Angela can only have Angela's attitudes. But it does not follow that Becca's attitudes are agent-relative in a normative way as well, and that they are relevant only for Becca, or that Angela's attitudes are agent-relative in a normative way and are relevant only for Angela. The claim that an agent's attitudes are agent-relative in the descriptive sense does not entail that they are also agent-relative in a normative sense.

To see this, consider the case of pain. Pain is also agent-relative in a descriptive sense. The existence of pain depends on the agent who experiences the pain, and *I* can only experience *my* pain, not yours, whereas *you* can only experience *your* pain, not mine. But that does not mean that my pain is normatively relevant only for me. My pain is bad, not only from my point of view but also in an agent-neutral sense (the badness of my pain does not depend on the fact that *I* am the one who experiences it), and other agents have reason to alleviate my pain if they are in a position to do so. Thus, the fact that my pain is agent-relative in a descriptive sense does not show that it is also agent-relative in a normative sense, and that it affects only my normative situation. Descriptive agent-relativity does not translate into normative agent-relativity. Accordingly, there is no rea-

---

nantia are all neutral values—that agents may have more reason to bring about those states of affairs." ("Love in the Time of Consequentialism," 10)

11 Maguire, "Values, Reasons, and Ought," 58.

son to think that the fact that my caring attitude cannot be anyone else's attitude but mine entails that my attitude only affects my normative situation, or that it modifies only the strength of my reasons.

II

One possible alternative is to ground the intensifying role of an agent's caring attitudes not in the value of those attitudes but in their actual occurrence, without saying anything about their value. In that case, it would not be obvious that the caring attitudes of an agent are normatively relevant for other agents—after all, it is the fact that the *value* of an agent's attitudes is doing the reason-modifying work that leads to the question of why the (agent-neutral) value of my attitudes does not modify your reasons as well. And it would also explain why the descriptive specialness of my attitudes translates into a normative specialness. The problem with this solution, however, is that it would rob the account of much of its plausibility. It would imply that an agent's actual caring about silly or evil outcomes modifies her reasons so that she possibly has most reason to perform silly or evil acts. It would also imply that Becca has more reason to enable Angela to save Angela's mother if, for some reason, Becca just does not care about the survival of her own mother.[12] Maguire's view is supposed to explain commonsense intuitions, and one of these intuitions is that Becca *does* have more reason to save her own mother, even if she does not care about her mother's survival. Every other claim would appear overly subjectivist.

Maguire's proposal faces a dilemma. Either he grounds the reason-modifying role of caring attitudes in their actual occurrence, rather than their value, in which case Becca has no more reason to save her own mother than to enable Angela to save Angela's mother if Becca simply does not care about the survival of her own mother. Or he grounds the reason-modifying role of an agent's caring attitudes in the value of those attitudes. He can then claim that Becca has more reason to save her own mother, even if she does not care about her mother's survival, because it would be good if she did care about her mother's survival. This, however, means that the value of a caring attitude can modify an agent's reasons, *even if that agent does not have that attitude.* This leads to the question of why the value of Angela's attitudes does not modify Becca's reasons also. Sure, Becca cannot have Angela's attitudes, but if the value of an attitude can modify an agent's reasons, even if the agent does not have that attitude, we need further explanation of why this does not also apply to the value of the attitudes of other

12 Maguire emphasizes these points; see "Love in the Time of Consequentialism," 10.

persons. As interesting as Maguire's proposal is, a convincing justification of partiality within a consequentialist framework is still pending.[13]

University of Bern
joerg.loeschke@philo.unibe.ch

REFERENCES

Bradley, Ben. "Against Satisficing Consequentialism." *Utilitas* 18, no. 2 (June 2006): 97–108.

Card, Robert F. "Consequentialism, Teleology, and the New Friendship Critique." *Pacific Philosophical Quarterly* 85, no. 2 (June 2004): 149–72.

Cullity, Garrett. "Neutral and Relative Value." In *The Oxford Handbook of Value Theory*, edited by Iwao Hirose and Jonas Olson, 96–116. Oxford: Oxford University Press, 2015.

Jollimore, Troy. *Friendship and Agent-Relative Morality*. New York: Garland Publishing, 2001.

Maguire, Barry. "Love in the Time of Consequentialism." *Noûs*. Published electronically August 25, 2016. doi:10.1111/nous.12169.

———. "Values, Reasons, and Ought." PhD diss., Princeton University, 2012.

McElwee, Brian. "Impartial Reasons, Moral Demands." *Ethical Theory and Moral Practice* 14, no. 4 (August 2011): 457–66.

McNaughton, David, and Piers Rawling. "Honoring and Promoting Value." *Ethics* 102, no. 4 (July 1992): 835–43.

Mulgan, Tim. "Slote's Satisficing Consequentialism." *Ratio* 6, no. 2 (December 1993): 121–34.

Portmore, Douglas W. *Commonsense Consequentialism*: *Wherein Morality Meets Rationality.* New York: Oxford University Press, 2011.

Schroeder, Mark. "Teleology, Agent-Relative Value, and 'Good.'" *Ethics* 117, no. 2 (January 2007): 265–95.

Slote, Michael. *Common-Sense Morality and Consequentialism*. London: Routledge, 1985.

———. "Satisficing Consequentialism." *Proceedings of the Aristotelian Society, Supplementary Volumes* 58 (1982): 139–64.

13 This work was supported by the Swiss National Science Foundation (SNSF) through the research project "The Normative Significance of Agent-Relative Reasons" (Grant 100012_152918). I would like to thank an anonymous referee for very helpful comments on an earlier version of this paper.

Smith, Michael. "Neutral and Relative Value After Moore," *Ethics* 113, no. 3 (April 2003): 576–98.

———. "Two Kinds of Consequentialism." *Philosophical Issues* 19 (October 2009): 257–72.

Vallentyne, Peter. "Against Maximizing Act-Consequentialism." In *Contemporary Debates in Moral Philosophy*, edited by James Dreier, 21–37. Malden: Blackwell Publishing, 2006.

Wallace, R. Jay. "Reasons, Values and Agent-Relativity," *Dialectica* 64, no. 4 (December 2010): 503–28.

Journal of Ethics and Social Philosophy
VOL. 12, NO. 1 · SEPTEMBER 2017
https://doi.org/10.26556/jesp.v12i1.218

AGAINST WOLTERSTORFF'S THEISTIC ATTEMPT TO GROUND HUMAN RIGHTS

David Redmond

NICHOLAS WOLTERSTORFF is concerned to find an appropriate grounding of human rights understood as inherent natural rights. Unfortunately, he thinks no adequate secular account of the grounding of such rights is available. Fortunately, however, he thinks that an adequate *theistic* account of the grounds of human rights is available. According to his proposed account, human rights are grounded in our standing in the relation of being loved by God.[1] After saying a word about how Wolterstorff understands rights in general and human rights in particular, I explain his proposed theistic account of the grounding of human rights and argue that it fails.

1. WOLTERSTORFF ON THE PROJECT OF GROUNDING HUMAN RIGHTS

For Wolterstorff, rights are a particular kind of normative social relationship. In particular, they are legitimate claims against another to the good of being treated in a way befitting of the right-holder's worth.[2] In order to possess a right, one must have a certain *status*. For example, to have the right to an NBA championship ring, one must have the status of having been a member of a team that won the NBA finals. A *human* right is a right such that the status sufficient for possessing the right is the status of being a human being.[3] That is, Wolterstorff assumes

1 See Wolterstorff, "Can Human Rights Survive Secularization." For a more complete explication of his argument, see Wolterstorff, *Justice: Rights and Wrongs*. He offers a revised account in *Justice in Love*, ch. 14. He offers yet another revision of the argument in "On Secular and Theistic Groundings of Human Rights." For his most recent defense, see "Why Naturalism Cannot Account for Natural Human Rights." Christopher J. Eberle offers a similar argument in "Basic Human Worth: Religious and Secular Perspectives."

2 For his detailed account of the nature of rights, see *Justice: Rights and Wrongs*, especially 241–310.

3 It is worth noting that Wolterstorff thinks of human rights—and, indeed, all rights—as *prima facie* rights. Even with this qualification, he thinks the common explanation is still not technically correct because he holds both that rights imply corresponding duties and that

what Tasioulas calls the "orthodox view": human rights are moral rights that one has simply in virtue of being a human being.[4] Many other contemporary writers agree with this characterization of human rights as natural rights. Griffin, for example, claims that human rights just are natural rights by a different name. He writes, "The French marked the secularization of the concept by changing its name from 'natural rights' to 'human rights' … The secularized notion that we were left with at the end of the Enlightenment is still our notion today. Its intension has not changed since then: *a right that we have simply in virtue of being human*."[5]

Wolterstorff recognizes that one can attempt to come to understand the nature of human rights either by starting from a standard list (like the UDHR) or by starting with the common explanation of human rights. The above makes it clear that, at least for present purposes, Wolterstorff has opted for the latter option. Those who favor the alternative starting place might worry that Wolterstorff does not correctly understand the nature of human rights.[6] And even those who are happy to begin with the common explanation of human rights might still suggest that the orthodox view must be qualified in some way or another.[7] For present purposes, such debates can be set aside. What is significant is that Wolterstorff takes as his starting point a widely accepted definition of human rights and claims that a secular grounding of human rights, so understood, is not available.

Though his exploration of human rights does not begin with the UDHR, he does agree with the UDHR in thinking that human rights are grounded in some dignity or worth that humans possess.[8] But as Wolterstorff points out, dignity "does not just settle on things willy-nilly."[9] There must be some feature about

---

there might arise circumstances in which no one has the relevant duty because no one is in a position to bestow the relevant good on the alleged right holder. In such a case, one's right is *blocked*. See "On Secular and Theistic Groundings of Human Rights," 179–81.

4 Tasioulas, "On the Foundations of Human Rights," 45.

5 Griffin, "On Human Rights," 1–2.

6 For an overview of the debate between naturalists and political conceptions of human rights, see Cruft, Liao, and Renzo, "The Philosophical Foundations of Human Rights: An Overview," 4–10.

7 See, for example, Griffin's explication of the term "human" in "human rights" in his "On Human Rights," 34–35. Cf. Simmons, "Human Rights, Natural Rights, and Human Dignity," 145. Wolterstorff admits that the writers of the UDHR seem not to be interested in human rights as he understands them; instead, they are concerned with what he calls "full-fledged human person rights," i.e., rights one has merely in virtue of being a full-fledged human person. See his "Grounding the Rights We Have as Human Persons," 203–4.

8 See his *Justice: Rights and Wrongs*, ch. 13, for a defense of a dignity-based grounding of rights.

9 Wolterstorff, "On Secular and Theistic Groundings of Human Rights," 182.

us on which the dignity supervenes. Identifying such a feature is the project of grounding human rights. The relevant feature must meet the following conditions: first, it is a feature that all human beings *must* have—a feature that no human can lack while remaining human. The reason for this is as follows: the status of being a human being is sufficient for possessing human rights. So, if human rights are grounded in dignity, then both the dignity that grounds human rights and the feature on which it supervenes must be inseparable from the status of being a human being. Second, because the entire package of human rights is unique to humans, the feature on which the dignity that grounds human rights supervenes must likewise be a uniquely human feature. And third, Wolterstorff and many others come to the table with the intuition that all humans have some dignity or worth in virtue of which they are due respect such that certain ways of treating them are unacceptable. The feature we're seeking must explain this; it must illuminate our intuition that all humans—even the significantly cognitively impaired—have a worth that explains why certain ways of treating them are unacceptable. Wolterstorff's contention is that there is no such feature (or combination of features) readily available to the secularist. He considers various secular accounts that attempt to ground human rights in the capacity for either rational agency or personhood. Unfortunately, such accounts clearly fail the first condition; not all humans have the relevant capacities. Furthermore, not only do some humans not *currently* possess such capacities, some humans never did and never will possess such capacities. Wolterstorff also considers a secular view according to which human rights are ultimately grounded in human nature. He argues, however, that while such a feature clearly meets the first two conditions, it fails to meet the third condition; possessing a human nature does nothing to illuminate our intuition that the significantly cognitively impaired have a worth that explains why certain ways of treating them are unacceptable. For purposes of this essay, I set aside whether Wolterstorff succeeds in defeating the above-mentioned secular accounts.[10] Instead, I focus on his proposed theistic grounding of human rights. In the remaining section, I explain his theistic account and argue that it fails.

2. Wolterstorff's Theistic Account of the Grounding of Human Rights

Perhaps the standard theistic account of the grounding of human rights appeals to the biblical notion of the *imago dei*. According to this account, humans have

10 Indeed, Wolterstorff would be pleased to discover that he has not. See his "Can Human Rights Survive Secularization," 420.

worth because they are made in the image of God. Of course, to evaluate such an account, we must first know how to understand the notion of the *imago dei*. Unfortunately, theologians have tended to understand it either directly in terms of the capacity for rational agency or personhood or else in terms that clearly imply those capacities.[11] If this is correct, then, if Wolterstorff's critique of secular accounts is successful, the theist cannot ground human rights in the *imago dei* because, once again, not all humans have the relevant capacities. Wolterstorff, therefore, offers an alternative theistic account according to which human rights are grounded in a particular relationship in which we stand to God—namely, the relation of being loved by God in such a way that worth is bestowed on us.

Not all species of love bestow worth, however.[12] For instance, one loves with "love as attraction" when the lover is attracted to the object of love in virtue of some aspect or worth of the object of love. "The love gives rise to desire for engagement with the things loved; the lover anticipates that his well-being will be enhanced by that engagement."[13] This sort of love will not do the work that Wolterstorff desires because it does not enhance or bestow worth; rather, it recognizes worth that is already present. By contrast, love as benevolence—at least when successful—does enhance worth, but "the enhancement does not consist simply in the person's being an object of benevolence; it consists in some alteration in his or her life that the lover causes."[14] For example, a man's life would go better if one provides him with food, but his receiving the food would have made his life better even if it was not motivated by benevolence.

In his original version of the argument, Wolterstorff argues that human rights are grounded in a third species of love, what he calls "love as attachment." He illustrates with a story of a boy named Nathan who loves his ugly stuffed animal.[15] Nathan even recognizes that it is ugly and that many other stuffed animals are nicer in various respects. Nevertheless, Nathan loves *his* stuffed animal. It is the one with which he has bonded; it is the one to which he is attached. This theistic account of the grounding of human rights was subject to extensive criticism,[16] and Wolterstorff no longer thinks that love as attachment can do the work he ascribed to it. He now concedes that God's loving all humans with love as attach-

11 Wolterstorff, "On Secular and Theistic Groundings of Human Rights," 194.

12 Wolterstorff distinguishes a variety of species of love in his *Justice: Rights and Wrongs*, 358–60.

13 Wolterstorff, *Justice: Rights and Wrongs*, 358–59.

14 Wolterstorff, *Justice: Rights and Wrongs*, 359.

15 Wolterstorff, *Justice: Rights and Wrongs*, 359.

16 See, for example, Bernstein, "Does He Pull It Off? A Theistic Grounding of Natural Inherent Human Rights?"; Weithman, "God's Velveteen Rabbit"; and Murphy, review of *Justice: Rights and Wrongs*.

ment merely gives *transitive moral significance* to each human. It explains why harming a human being may be a way of wronging God, but it does not explain why harming a human being is a way of wronging the human being.[17]

In his most recent treatments of the topic, Wolterstorff argues that there is yet another kind of love that can do the trick. He calls it "love as the desire for friendship."[18] He invites us to imagine a good and beloved monarch who, because he is lonely, decides to befriend a few of his subjects. Those chosen are honored to be chosen as his friends. They "will cite that fact under 'Honors' in their *curricula vitae*."[19] He continues:

> Now for the crucial point. To be honored is to have worth bestowed on one. Admittedly, this is somewhat mysterious. But an indication of the fact that this is what happens is that to be honored is to acquire new grounds for respect and, hence, new ways in which one can be demeaned and wronged; and that's possible only if there has been some alteration in one's worth. The most obvious way in which one can now be demeaned and wronged is by having the honor itself belittled. One who has not been chosen by the monarch for friendship sarcastically remarks to someone who has been chosen, "Big deal!" The latter has a right to be angry; this is a snub. Depending on the situation, the snub may count as a snubbing of the one who bestowed the honor; but in any case, it amounts to snubbing the one honored. Honoring bestows worth.[20]

The relevant analogy is obvious. A theist might plausibly maintain that God desires friendship with all human beings. Indeed, many Christians believe that the metanarrative of all history involves God's pursuit of a restored relationship with each of the humans he created. Such a desire for human friendship constitutes a way in which God honors us, and being honored by God in this way is to have worth bestowed on us. Furthermore, God's desire for friendship with human beings rather than, say, crocodiles is not arbitrary because, unlike crocodiles, we have the potential for friendship with God because it is in our nature to be persons.

Jordan Wessling has recently argued that Wolterstorff's account is subject to something similar to the traditional Euthyphro dilemma.[21] If he is right, then we

17 Wolterstorff, "On Secular and Theistic Groundings of Human Rights," 196–97.

18 See Wolterstorff, "On Secular and Theistic Groundings of Human Rights," 197–200, and "Why Naturalism Cannot Account for Natural Human Rights," 259–61.

19 Wolterstorff, "On Secular and Theistic Groundings of Human Rights," 198.

20 Wolterstorff, "On Secular and Theistic Groundings of Human Rights," 198.

21 Wessling, "A Dilemma for Wolterstorff's Theistic Grounding of Human Dignity and Rights."

have a rebutting defeater for Wolterstorff's account. But regardless of whether Wessling is right, Wolterstorff's argument is subject to an undercutting defeater. As he presents the argument, it depends on two claims: (1) The monarch's desiring friendship with his subjects is a way of honoring them, and (2) Honoring his subjects in this way bestows worth. Suppose we grant to Wolterstorff the first claim.[22] Still, the crucial point concerns whether we have any reason to think that being honored in this way *bestows* worth. Wolterstorff admits that this is mysterious. I agree, and I think I can explain the mystery: his argument that honoring bestows worth is spurious.

Wolterstorff claims that those subjects with whom the monarch desires to be friends has acquired new ways in which he or she can be demeaned and wronged. Call one such subject, "James." He writes, "The most obvious way in which [James] can now be demeaned and wronged is by having the honor itself belittled. One who has not been chosen by the monarch for friendship sarcastically remarks to [James], 'Big deal!' … This is a snub."[23] We should ask two questions about James: (1) Does James acquire new ways in which he can be wronged in virtue of the monarch's desire for friendship? (2) If so, does this indicate that the monarch's desire bestowed worth on James? The success of Wolterstorff's arguments depends on his defense of an affirmative answer to both questions. Unfortunately, his defense fails. To begin with, it is not clear that James does acquire new ways of being wronged.[24] To see this, imagine that the monarch does not, in fact, desire friendship with James, but both James and his acquaintance falsely believe that he does. In this case, it seems that James would be wronged in precisely the same way if his acquaintance exclaimed, "Big Deal!" This suggests that James does not acquire new ways of being wronged in the case that Wolterstorff imagines.

But suppose I am wrong about that. Suppose James does acquire new ways in which he can be wronged. In this case, however, James's acquiring new ways in which he can be wronged is not indicative of new worth. To see this, return to the original case in which the monarch really does desire friendship with James, and his acquaintance sarcastically responds, "Big deal!" It is plausible that such a response constitutes wronging James insofar as it is disrespectful to show

22 Because I am not sure what exactly is involved in honoring someone, I am unsure whether the monarch's desiring friendship with some individual, *X*, constitutes honoring *X*. In any case, the premise is superfluous. Wolterstorff's argument succeeds if he can establish both that God's desire for friendship with humans generates new ways in which they can be demeaned and wronged and that the generation of such ways is indicative of new worth.

23 Wolterstorff, "On Secular and Theistic Groundings of Human Rights," 198.

24 This depends, in part, on how one individuates ways of being wronged.

contempt for his experiencing some good in his life. We need not think that it is due to some new worth that was bestowed on him.[25] Modify the case once again. Imagine, instead, that James has won a new car and is excited about his vehicle. He would be wronged if his acquaintance responded, "Big deal!" The explanation I suspect is the same in both cases; in neither case do we need to suppose that James is wronged because of some new worth that he acquired. Rather, James is wronged because someone showed disrespect for a person (who already possesses worth) by showing contempt for the acquisition of some good in his life. So, even if we grant that James acquires a new way in which he can be wronged when the monarch comes to desire friendship with him, this does not imply that he has undergone an alteration in worth. One can acquire new ways of being wronged just in virtue of acquiring new goods in one's life. If that is correct, then Wolterstorff fails to adequately defend his theistic account of the grounding of human rights. While he may have located a feature that—given some plausible theistic assumptions—meets his first and second conditions, he provides no reason to suppose it meets the third. His attempt to argue that it does is spurious because not only is it unclear whether God's desire for friendship with humans generates new ways in which they can be wronged, but even granting that it does, acquiring new ways in which one can be wronged does not indicate that worth has been bestowed. Without justification for thinking that worth has been bestowed, Wolterstorff provides no justification for thinking that the theist succeeds where the secularist—at least in Wolterstorff's estimation—fails. If all humans have a dignity or worth in virtue of which they have certain rights, then, if Wolterstorff's criticisms of the secular accounts are cogent, a dignity-based grounding of those rights remains elusive.[26]

University of Iowa
david-redmond@uiowa.edu

25 Wolterstorff recognizes that, because a human is a substance and his or her life is an event, a human being is not identical with his or her life. He also explicitly notes that, "from the fact that the state of affairs of *my being K* is a good in my life, it does not follow that my having the property of *being K* makes a positive contribution to my worth as a human being." See *Justice: Rights and Wrongs*, 142–43. It seems to me that Wolterstorff's revised theistic grounding of human rights neglects this important fact.

26 Many thanks to the members of the University of Iowa Graduate Philosophical Society as well as the participants at the 2016 Midsouth Philosophy Conference, the 2016 Midwest meeting of the Society of Christian Philosophers, and the 2017 Jakobsen Memorial Conference for discussions on earlier drafts.

REFERENCES

Bernstein, Richard J. "Does He Pull It Off? A Theistic Grounding of Natural Inherent Human Rights?" *Journal of Religious Ethics* 37, no. 2 (June 2009): 221–42.

Cruft, Rowan, S. Matthew Liao, and Massimo Renzo, eds. *Philosophical Foundations of Human Rights*. Oxford: Oxford University Press, 2015.

———. "The Philosophical Foundations of Human Rights: An Overview." In Cruft, Liao, and Renzo, *Philosophical Foundations of Human Rights*, 1–44.

Eberle, Christopher J. "Basic Human Worth: Religious and Secular Perspectives." In *New Waves in Philosophy of Religion*, edited by Yujin Nagasawa and Erik J. Wielenberg, 167–91. New York: Palgrave Macmillan, 2009.

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

Murphy, Mark. Review of *Justice: Rights and Wrongs*, by Nicholas Wolterstorff. *Ethics* 119, no. 2 (2009): 402–7.

Simmons, A. John. "Human Rights, Natural Rights, and Human Dignity." In Cruft, Liao, and Renzo, *Philosophical Foundations of Human Rights*, 138–52.

Tasioulas, John. "On the Foundations of Human Rights." In Cruft, Liao, and Renzo, *Philosophical Foundations of Human Rights*, 45–70.

Weithman, Paul. "God's Velveteen Rabbit." *Journal of Religious Ethics* 37, no. 2 (June 2009): 243–60.

Wessling, Jordan. "A Dilemma for Wolterstorff's Theistic Grounding of Human Dignity and Rights." *International Journal for Philosophy of Religion* 76, no. 3 (December 2014): 277–95.

Wolterstorff, Nicholas. "Can Human Rights Survive Secularization." *Villanova Law Review* 54, no. 3 (2009): 411–20.

———. "Grounding the Rights We Have as Human Persons." In *Understanding Liberal Democracy: Essays in Political Philosophy*, 201–26.

———. *Justice in Love*. Grand Rapids, MI: Eerdmans, 2011.

———. *Justice: Rights and Wrongs*. Princeton, NJ: Princeton University Press, 2008.

———. "On Secular and Theistic Groundings of Human Rights." In *Understanding Liberal Democracy: Essays in Political Philosophy*, 177–200.

———. *Understanding Liberal Democracy: Essays in Political Philosophy*. Edited by Terence Cuneo. Oxford: Oxford University Press, 2012

———. "Why Naturalism Cannot Account for Natural Human Rights." In *The Blackwell Companion to Naturalism*, edited by Kelly James Clark, 447–61. Malden, MA: Wiley-Blackwell, 2016.