JOURNAL *of* ETHICS
& SOCIAL PHILOSOPHY

VOLUME XXVII · NUMBER 1
March 2024

ARTICLES

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.

Editors
Sarah Paul
Matthew Silverstein

Associate Editors
Rima Basu
Saba Bazargan-Forward
Brian Berkey
Ben Bramble
Dale Dorsey
James Dreier
Julia Driver
Anca Gheaus
Alex Gregory
Sean Ingham
Anthony Laden
Errol Lord
Coleen Macnamara
Elinor Mason
Simon Căbulea May
Hille Paakkunainen
David Plunkett
Mark van Roojen
Kevin Toh
Vanessa Wills

Discussion Notes Editor
Tristram McPherson

Managing Editor
Chico Park

Copy Editor
Lisa Gourd

Proofreader
Susan Wampler

Typesetter
Matthew Silverstein

Editorial Board
Elizabeth Anderson
David Brink
John Broome
Joshua Cohen
Jonathan Dancy
John Finnis
Leslie Green
Karen Jones
Frances Kamm
Will Kymlicka
Matthew Liao
Kasper Lippert-Rasmussen
Stephen Perry
Philip Pettit
Gerald Postema
Henry Richardson
Thomas M. Scanlon
Tamar Schapiro
David Schmidtz
Russ Shafer-Landau
Tommie Shelby
Sarah Stroud
Valerie Tiberius
Peter Vallentyne
Gary Watson
Kit Wellman
Susan Wolf

Journal of Ethics and Social Philosophy
vol. 27, no. 1 · march 2024

https://doi.org/10.26556/jesp.v27i1.2843
© 2024 Author

REMAINING TRUE TO OURSELVES

DEMENTIA, VALUE CHANGE, AND ENDURING INTERESTS

Andrew Franklin-Hall

It is common to think that when we need to make choices in the best interests of others, we should be guided (at least in part) by their values.[1] But the later stages of dementia may cause people to lose a grip on some of the things that used to be important to them. Should we then consider their former values, even though we do not ordinarily consider what people *used* to care about in trying to discern their best interests? Or should we simply be guided by the person's perspective as it is now (and as it will be in the future)? Consider the following cases:

> *Mr. Amato*: For the last two years, Mr. Amato, who has Alzheimer's disease, has been living in a long-term care facility. He has been reasonably content there, making friends with other residents and enjoying group activities. However, it has recently come to the attention of his children, John and Sally, that their father has been walking around common areas half-disrobed and that he is seldom afforded any real privacy in his room. When they asked the administrators about this, they were told that since Mr. Amato does not seem to mind these things, it is easiest not to worry about them. John and Sally remain disturbed, though, because they know that their father used to consider propriety and privacy as integral to living in a dignified way.

> *Ms. Bell*: Although Ms. Bell has never received a formal diagnosis, she is plainly living with relatively advanced dementia. Nevertheless, with the help of home health aides, she has been able to remain in her house. Recently, her daughter, Claire, noticed certain family heirlooms missing from her mother's shelves. Suspecting the caregivers of theft at first, she eventually discovered that her mother had been giving these things away

[1] See, for example, Buchanan and Brock, *Deciding for Others,* 29–36.

> to neighbors and friends (and taking real joy in doing so). Claire had mixed feelings about this. She did not relish inheriting all those objects, but she also knew that it used to mean a lot to her mother to keep them in the family. When Claire tries to remind her mother of this, Ms. Bell tends to make a joke and change the subject.

It is not my aim to prescribe exactly what we ought to do in cases like these. Such decisions will call for considerable judgment, empathy, and creative problem-solving. My concern is with the more fundamental question as to whether the former values of people in circumstances like those just described have any bearing on what is in their best interests—or, put differently, is prudentially valuable for them or contributes to their "good," "welfare," or "well-being." These terms have different shades of meaning in ordinary language, but like many philosophers today, I use them synonymously. I do not know how to define well-being in any more basic terms, but we do seem to use some such notion in thinking about a web of related concepts, like benefit, advantage, harm, and enlightened self-interest. Crucially, as most philosophers today understand the notion, it is not an analytic truth that well-being is equivalent to subjective quality of life. For example, while there is disagreement about whether things external to a person's consciousness can affect her well-being, most philosophers regard that as a coherent claim.

The value of autonomy tends to loom large in discussions about decision-making for patients lacking decision-making competence. Insofar as autonomy or self-determination is conceived of as one thing that is good for a person, it is part of my topic.[2] However, there is also a purely deontic notion of personal sovereignty over one's life that, to keep the discussion manageable, I will not address. Likewise, I set aside here obviously relevant questions about how these decisions might affect the interests of others. This means that I am not offering an account of what we should do, all things considered, in the cases under consideration.

Two approaches to our problem have been especially prominent in the philosophical literature on dementia. On what we might call the *Presentist Model,* what a person used to value simply cannot bear on her current interests. To have an interest in something, they think, one must at least remain capable of caring about it or taking an interest in it. Thus, Rebecca Dresser holds that

> individual incompetent patients' interests are invariably a function of their physical and mental capacities.... A truly patient-centered best interests assessment will incorporate an examination of the particular incompetent patient's interests in light of his individual capacities....

[2] See, for example, Griffin, *Well-Being,* ch. 4; and Raz, *The Morality of Freedom,* ch. 14.

> Matters such as dignity, privacy, and bodily integrity arguably are integral to the well-being of the average or reasonable competent person in our culture. But it is nonsense to claim that these matters affect the well-being of many incompetent patients with severely compromised mental abilities.[3]

An alternative view says that the capacity to value, or care, or find meaning in things bestows a certain standing on a person's present perspective, and this makes it inappropriate to appeal to her former values *until* the person no longer has this capacity. At that point, however, a person's former values can matter. I will refer to this as the *Status Model*.

I side with the Status Model in thinking that there are conditions in which a person's past values can bear on his current interests. But I deny that we need to attribute exclusive authority to either the person's present perspective or his past one. Instead, I defend a more piecemeal approach. A person's past values can matter even while he continues to value or care about simpler things now. The critical question, as I see it, is whether a person is answerable for the fact that he no longer holds a certain value that was once dear to him. For that, he must still be able to understand what he used to value and why—he must, in other words, retain the capacity to critically reflect on that past value. If he cannot, we should not think of him as having revised his former value. Rather, that value remains imputable to him as his last authentic verdict about the matter. Call this the *Revision Model*. To defend this view, I will begin with a discussion of a few influential versions of the Status Model, then outline the Revision Model, and finally explain why I think it is more plausible than the Presentist Model.

Often this issue is discussed within medical ethics as a part of a conversation about advance directives, substitute decision-making, and end-of-life care. This makes sense in that the stakes are so high in those cases. But for that same reason, I believe those cases can have a distorting effect on our analysis, for they tend to be represented in stark either/or terms, and they involve questions of life and death, which may raise additional moral issues and pertinent questions of public policy. For this reason, they may distract us from the fundamental question as to whether a person's former values can still have some bearing on his interests. That is why I will be focusing on more mundane cases, like those of Mr. Amato and Ms. Bell. My conclusions should be relevant to the more familiar debates in medical ethics, but they will not settle them.

Finally, I should acknowledge that some have thought that the deterioration of memory and personality can be so profound that the dementia patient is no longer the same person, metaphysically, as she used to be. In that case, it would

[3] Dresser, "Life, Death, and Incompetent Patients," 383–85.

seem wrong to impose the former person's values on the present individual.[4] This relies on a controversial theory of personal identity, but pursuing that issue would take us too far afield.[5] I will only note that, even accepting that view, there will be cases where someone with dementia has lost hold of a former value, and yet enough of his memory and personality remain that it would be hard to deny this is still the same person. So, appeal to personal identity will not entirely dispose of the puzzle.[6]

1. WELL-BEING AND THE STATUS MODEL

I shall take, as my point of departure, the influential views of Ronald Dworkin and Agnieszka Jaworska.[7] Although their accounts differ in their practical implications, they both defend versions of what I am calling the Status Model.[8] I begin with Dworkin's ideas about the objective and subjective dimensions of our so-called critical interests, then outline the general structure of the Status Model, and finally raise some difficulties for it.

1.1. The Objective and Subjective Aspects of Well-Being

According to Dworkin, although we all have "experiential interests," which involve having agreeable experiences and avoiding disagreeable ones, our more important decisions in life are typically guided by our sense of our "critical interests." These critical interests have both an objective and subjective dimension, according to Dworkin.[9] From the agent's point of view, critical interests look largely objective. Most people, he says, take themselves to have "interests that … make their life genuinely better to satisfy, interests they would be

[4] Dresser, "Life, Death, and Incompetent Patients" and "Dworkin on Dementia." For criticism of this idea, see Buchanan and Brock, *Deciding for Others*, ch. 3; and DeGrazia, *Human Identity and Bioethics*, ch. 5.

[5] Alternative views are defended in McMahan, *The Ethics of Killing;* and DeGrazia, *Human Identity and Bioethics*.

[6] Nor do I discuss "time-relative interests" (see McMahan, *The Ethics of Killing*). That past values can matter is consistent with thinking their importance should be discounted the less psychologically connected the person is to his past self.

[7] Again, I focus on arguments appealing to well-being, not autonomy.

[8] Called the "threshold of authority approach" in Jaworska's entry on "Advance Directives and Substitute Decision-Making" in the *Stanford Encyclopedia of Philosophy*.

[9] Dworkin, *Life's Dominion*, 206. My interpretation of critical interests is informed not only by Dworkin's discussion of dementia and end of life in *Life's Dominion*, but also by his more systematic treatment in "Foundations of Liberal Equality," to which we are referred in an endnote in *Life's Dominion* (255n21). (The relevant sections of "Foundations" are also reprinted, with minor amendments, as chapter 6 of *Sovereign Virtue*.)

mistaken, and genuinely worse off, if they did not recognize."[10] Hence, our critical interests involve the kinds of things we commonly find on "objective list" theories of well-being: instances of knowledge, achievement, living up to worthy ideals, maintaining valuable relationships, and so on.[11] These are things that we conceive of ourselves as wanting because they are valuable—things we admire in the lives of others and that we can reflect on in our own lives with pride and satisfaction.

But critical interests also have a subjective side, especially from the observer's point of view. Part of Dworkin's idea is that what makes a person's life go better is the way something fits in with the distinctive character of that particular life.[12] This tends to privilege the values that a person has been living by to this point. But, for a couple of reasons, Dworkin also thinks of a person's current perspective as ordinarily bearing preeminent authority in defining his current critical interests.

First, Dworkin subscribes to the "endorsement thesis": that nothing can noninstrumentally contribute to our well-being (in the critical sense) unless we endorse its value at some level.[13] "Value," as he says, "cannot be poured into a life from the outside; it must be generated by the person whose life it is."[14] This leads him to assert that a person's critical interests are partly constituted by his convictions.[15] The endorsement thesis does not imply that a person will benefit from something *merely* because he endorses it; the object must actually be valuable. But, presumably, most of what we value is not entirely worthless. So, my life might have been better if I valued family more than achievement, but if I only appreciate achievement to any significant degree, then (holding

[10] Dworkin, *Life's Dominion*: 201–2. Compare Dworkin, "Foundations of Liberal Equality," 230, and *Sovereign Virtue*, 245.

[11] Dworkin, *Life's Dominion,* 201–2. On the objective-list theory, see Parfit, *Reasons and Persons*, 499–501.

[12] Dworkin, *Life's Dominion,* 202, 205–6. Compare Dworkin, "Foundations of Liberal Equality," 249–53, and *Sovereign Virtue*, 257–60.

[13] Dworkin, "Foundations of Liberal Equality," 237, 264, and *Sovereign Virtue*, 248–49, 268. In *Life's Dominion,* Dworkin writes: "It is important both that we find a life *good* and that we *find* it good" (206).

[14] Dworkin, *Life's Dominion*, 230.

[15] To avoid circularity, Dworkin probably should have distinguished between what would be a critical interest for me if I endorsed it ("presumptive critical interests") and what is good for me given what I actually endorse ("confirmed critical interests"). Our convictions are partly constitutive of the latter, not the former. When I deliberate about my own life, I mainly focus on my presumptive critical interests, but benefactors need to take my confirmed critical interests into account.

my values fixed) what will actually contribute to my critical interests is achievement, not family.

The other idea is that there is independent worth in my figuring out for myself what is right for my life and putting that vision into effect. This is at least part of what Dworkin means by "integrity."[16] The good of living with integrity lies not in the content of my values (which may be misguided) but in the way that I am faithful to my understanding of what is valuable.[17]

I am sympathetic to this general picture. I think we do typically assume an objective stance when we reason about how to live.[18] And yet, it also seems that a person's interests (at least from the bystander's perspective) are partly defined by her own values and convictions. Indeed, I am inclined to think that Dworkin is basically right that a person cannot noninstrumentally benefit from a good (at least not substantially) unless it finds some purchase within his evaluative perspective and that there is significant value in living a life according to one's own convictions. But what is most important to my argument going forward is that Dworkin is right to think that our more sophisticated interests have both an objective and subjective side.

1.2. The Status Model

If a person's well-being partly depends on his own ideas about the good life, then we must consider those when trying to promote his interests. But what if a person does not have such views anymore? Dworkin maintains that when a person loses his "sense of life as a whole" and thus no longer has a coherent view of what living a good life consists in, we should look back to the views he formerly held.[19]

[16] "Someone has ethical integrity . . . when he lives out of the conviction that . . . no other life he might live would be a plainly better response to the parameters of his ethical situation rightly judged (Dworkin, *Sovereign Virtue*, 270; compare "Foundations of Liberal Equality," 267). In *Life's Dominion*, Dworkin adds the diachronic idea that a life ought to "display a steady, self-defining commitment to a vision of character or achievement that the life as a whole, seen as an integral creative narrative illustrates and expresses (205). I think Dworkin's considered view is that my life has integrity when I live consistently with my current view about what is fitting for my life, given the way I have lived it to this point.

[17] In *Life's Dominion*, Dworkin ties the right to autonomy to its protection of the agent's *capacity* for integrity, whether integrity is actually achieved or not (224). But the present claim is that *achieving* integrity actually makes a person's life go better. In fact, Dworkin says that "a life that never achieves . . . integrity cannot be critically better for someone to lead than a life that does" (*Sovereign Virtue*, 270; compare "Foundations of Liberal Equality," 267). See also the remarks in *Life's Dominion* about the value of a life partly consisting in its achieving integrity (206, 224).

[18] I take no position on the metaethics of this objective stance.

[19] Dworkin, *Life's Dominion*, 230.

To see the structure of the position, let us say that an attitude has *prudential authority* if something is noninstrumentally good for the person, at least in part because the attitude favors that thing. So, according to a simple desire-satisfaction theory, a person's desires have prudential authority. On Dworkin's view, the attitude with prudential authority is a certain kind of judgment about what makes one's life good or successful. By contrast, purely objective theories of well-being deny that any subjective attitude has prudential authority.

Dworkin's suggestion is that, as long as a person continues to express attitudes possessing prudential authority, the person's current perspective has the exclusive standing to pick out what is valuable for her life (insofar as this depends on her subjective attitudes). This is because, as Dworkin says, such a person retains "the ability to act out of a genuine preference or character or conviction or sense of self," and we have an interest in living in accordance with our convictions.[20] Call this the *Status Model*. Expressing attitudes with prudential authority confers on a person's present outlook a standing or status that makes it inappropriate to appeal to that person's former perspective in working out what is best for her.[21] Plainly, you could accept the Status Model but differ with Dworkin over which attitudes possess prudential authority.[22] After all, to say that one needs convictions about what is good for one's life as a whole is a demanding view. If simpler attitudes could bear prudential authority, then the Status Model would imply that a person's present outlook could retain the exclusive standing to define what is best for her until a later stage of cognitive deterioration.

This is the key move Jaworska makes. In her 1999 article, she argues that it is a person's values that possess prudential authority. To value something, she says, just involves being able to give some account of what makes it good or worthwhile and some recognition that such judgments are, in principle, open to revision in the light of better reasons. Valuing, however, need not involve global conceptions about what makes a life successful; it can just consist in local judgments about what seems right for us, here and now. But these local values still inform our sense of ourselves as agents answerable to certain standards and typically affect our sense of pride or shame. Thus, even simple values enable a person to act (as Dworkin might say) from a "sense of self."[23] The implication is that many people living with mid-stage Alzheimer's disease remain valuers, even if they are not capable of the kind of whole-life assessment that Dworkin

[20] Dworkin, *Life's Dominion*, 225.

[21] This resembles the concept of "full moral standing" in Jaworska, "Caring and Full Moral Standing."

[22] You could also demur from Dworkin's commitment to the objectivity of our interests.

[23] For independent confirmation of this claim, see the conversations transcribed in Sabat, *The Experience of Alzheimer's Disease*.

privileges. Someone with this kind of impairment should, Jaworska thinks, "be viewed as any other person whose values and commitments change over time and whose currently professed values are taken to bear on what is best for her."[24] Only when a person has lost the ability to value in this modest way should we resort to her former values.

In subsequent work, Jaworska sets the bar even lower.[25] The right threshold for exclusive prudential standing, she claims, is the capacity to *care*, where this involves a significantly intense, sustained, and cohesive pattern of emotional vulnerability and attentiveness to something, which construes the object of concern as important to the agent. Although valuing something typically involves caring about it, caring is a simpler attitude since it need not involve judgments about what is genuinely "good" or "right," nor need it be associated with self-reflective attitudes (like pride or shame). But it is this capacity for sustained emotional attunement that, according to Jaworska, underlies the possibility of a cohesive and enduring self. Therefore, only a self of this sort can have attitudes with the kind of prudential authority that expresses a "sense of self." It is only once a person loses this more basic capacity to care that we should consult his former values since only at that point is there no longer a "deep self" present.[26] For instance, even if Ms. Bell has lost some of her past values, she evidently continues to value or care about making others happy, and that means that her current perspective retains the exclusive authority in defining her interests.

Jaworska makes a strong case that the attitudes with prudential authority need not be as sophisticated as Dworkin suggests. In fact, you might think that even Jaworska's view is too demanding. If the intuition behind the endorsement thesis is that I cannot directly benefit from an objective good unless it resonates with me, that it cannot be alien to me, then why is it not enough that I appreciate the good in the moment? Why does this receptivity have to be a stable disposition (like valuing or caring)? Perhaps the idea is that critical interests are supposed to make our *lives* better, and for this, they must structure some significant portion of our lives. Perhaps only temporally extended attitudes can support the kind of investment in things that makes this possible. If that is how we define critical interests, though, it seems we should acknowledge that we also have more episodic ways of benefiting from objective goods that, nonetheless, are not classifiable as mere experiential interests.

[24] Jaworska, "Respecting the Margins of Agency," 112.

[25] Jaworska, "Caring and Full Moral Standing." See also Shiffrin, "Autonomy, Beneficence, and the Permanently Demented," 211.

[26] Should we, at that point, appeal to the values the person held prior to impairment or to the last set of genuine concerns held while impaired? The former view is more intuitive, but the latter seems more consistent with the logic of the position.

1.3. Problems with the Status Model

That said, I think that some difficulties arise for even the best version of the Status Model. First, although Dworkin and Jaworska tell us *when* they think we ought to resort to a person's former attitudes, they do not really explain why our former values matter in just this way. If a person's past attitudes have any validity, why should they not normally be counted—even for people without impairments?[27] On the other hand, if they do not normally count, then why suppose that they *can* play this reserve role of coming forward just when the person loses the ability to express attitudes with prudential authority?

Second, suppose that, following Jaworska, we adopt an expansive view of the attitudes that bear prudential authority. Perhaps Ms. Henderson had long valued tithing to her church and intended to continue to give for the remainder of her life. Now, however, her dementia has deprived her of any understanding of that former commitment. The Status Approach says that whether Ms. Henderson still has an interest rooted in that former value depends on whether she still has prudentially authoritative values or concerns for anything else. But why should the moral relevance of this former commitment depend on whether the subject continues to care about other, unrelated matters? Indeed, what if making those donations on behalf of Ms. Henderson would not interfere with anything that presently matters to her? It seems like whether a former value constitutes an enduring interest should turn on the person's relation to that very value.[28]

To see the third issue, return to the case of Ms. Bell. Suppose that, at age eighty-five, she ceased to care about passing on family heirlooms to her children. On Jaworska's caring version of the Status Model, since Ms. Bell continues to care about other things, that former value is irrelevant to her best interests. But suppose that, by the time she turns eighty-eight, she really cannot be said to "care" about anything (in Jaworska's sense). At that point, apparently, she does have an interest in passing on the heirlooms again. But it seems strange to think that the relevance of this value to Ms. Bell's interests should fade away and then return in this manner. Stranger still, when Ms. Bell is eighty-five, we can predict that, while passing on the heirlooms is not currently good for her, it will be good for her later (if she lives long enough). But if acting prudently for Ms. Bell

[27] Dworkin might appeal to the importance of a life remaining consistent with the themes that have subsequently structured it, but then we have grounds for overriding a fully competent person's decision to keep him faithful to the sort of life he has already been leading (see Hawkins, "Well-Being, Time, and Dementia," 525–26).

[28] A slightly different view would be that past values can matter as long as they do not directly conflict with a person's present values or more robust concerns. I would also reject this approach, however, because I doubt that the slightest current value always trumps the weightiest former value.

means taking her current and (probable) future interests into account, should we not already (while she is eighty-five) think she has a future-oriented interest in passing on the heirlooms? It seems more plausible to either say that this value has always mattered during the period of Ms. Bell's impairment or that it never matters. In this light, Jaworska's version of the Status Model starts to look like an uneasy compromise between those options. These problems do not arise so acutely on Dworkin's version of the Status Model since it is quicker to resort to a person's former values. But this advantage is bought at the unacceptably high price of ignoring the critical interests rooted in Ms. Bell's present perspective.

2. THE REVISION MODEL I: THE PRIORITY OF VALUING

2.1. The Revision Model: An Overview

I call the view I defend the *Revision Model*, though it might more properly be labeled the "Authentic" or "Authoritative Revision Model." Here is the basic idea. Although a range of subjective attitudes of varying complexity can bear prudential authority, a person's values have priority in this respect since these represent the agent's own considered—and thus authentic or authoritative—perspective about his interests. Ordinarily, we should attribute the highest prudential authority to a person's *current* values (not his former ones) because these represent his most up-to-date revision of his earlier views. However, if a person has lost the ability to comprehend some of his former values, then we cannot treat his current perspective as a genuine revision since he is not answerable for the fact that his values have changed in this respect. Having never truly revised those former values, we should recognize their enduring authority in defining his interests insofar as they remain relevant. However, this is consistent with recognizing that the person may, at the same time, have other interests rooted in his current perspective. In this section, I make the case for the priority of values and emphasize the rational element in value revision. In section 3, I address the authority of a person's values at different points in time. And in section 4, I consider the relative importance of different sorts of past values for a dementia patient's current interests, as well as the significance of conflicts between a person's present and past perspectives.

2.2. Prudential Authority and Agential Authority

Any view recognizing the prudential authority of subjective attitudes must explain how to handle the relation between a person's various concurrent attitudes, especially when these conflict with one another. One approach would be to assign weights to attitudes in proportion to their phenomenal intensity. But this is not how we reflect on our own feelings and desires. That I feel a strong urge to do something does not at all settle for me, on reflection, that this desire is important to satisfy, for I might entirely repudiate some of my feelings and desires. If we simply attribute authority to desires or other attitudes based on their felt intensity, then we are treating the person not as an active agent but as a mere site of competing emotions and motivations. To take agency seriously, we must consider the person's own view about the relative importance of her various attitudes (assuming she has such a view). But to do this, we need to characterize the perspective that counts as the agent's authentic standpoint: the point of view from which she authoritatively identifies with some of her attitudes, rejects others, and settles on what weight to attribute to her various priorities. And, if we do not want to posit a little pilot inside the head who is the "true self," we need to identify a person's authentic perspective with certain attitudes, or configurations of attitudes, under certain conditions. Let us say that the attitudes that comprise a person's authentic perspective possess "agential authority."[29] My first claim, then, is that prudential authority follows agential authority. That is, the attitudes that constitute a person's authentic perspective as an agent have the greatest authority in defining what the person's true interests are.[30]

At a verbal level, many philosophers would agree that, for the mature, unimpaired agent, this authentic perspective is framed by what he, on reflection, and in a calm and collected state, values.[31] Valuing, here, should be understood in a dispositional way. I do not lose my values when I am asleep or when they are not present before my mind. To speak of "a collected state" means we are especially interested in what he values when he can reflect on how his different interests relate and compare. So, this suggests that Dworkin is on the right track in thinking our more global values tend to have a certain priority for those of us who have them. The main point of controversy lies not in whether values are important but in how we think about the kind of attitude valuing is.

[29] See, for instance, Bratman, "Reflection, Planning, and Temporally Extended Agency" and "A Desire of One's Own." The kind of "authentic perspective" or "agential authority" that Bratman identifies emerges from the influential work of Frankfurt ("Freedom of the Will and the Concept of a Person") and Watson ("Free Agency"). Similarly, Brudney and Lantos speak of living in accordance with one's commitments as the value of "authenticity" ("Agency and Authenticity").

[30] Compare Raibley, "Well-Being and the Priority of Values"; and Dorsey, *A Theory of Prudence*, ch. 5.

[31] There may be other conditions, such as freedom from manipulation or oppressive socialization, but I leave that aside.

Before expanding on that, though, let me clarify: I do *not* say that nothing can be good for us unless we value it. There are, as I have indicated, simpler attitudes that can open us up to the good things in life and support our investment in them. People with impairments may not be capable of the kind of reflection involved in valuing. And none of us reflects on all our attitudes. What is special about valuing is that it can overrule and constrain the presumptive authority of our other attitudes. I might find myself caring a lot about recognition for my accomplishments, and yet, on consideration, decide that I have been much too focused on this. In this way, my values represent my considered view, insofar as I have one, about what is important in my life and how important it is.[32] So, to the extent that my subjective attitudes have some bearing on what my interests are, my values simply have the highest authority. That said, normal adults typically *have* reflected to some degree on the central ideals, projects, activities, and relationships that structure their lives. Hence, for them, their key interests usually *are* defined by their values since these represent their own perspectives about what is important in their lives.[33]

2.3. The Nature of Valuing

Although many would agree that the agent's authentic perspective is defined by his values, there are different views about what this attitude consists in. The accounts we get from Dworkin and Jaworska are broadly cognitivist in spirit in that judgments of value are conceived to be central to the nature of valuing. Dworkin says that when we are deliberating about how to live, we are trying to form a judgment about "what is really important in life," which is something we think we can be mistaken about.[34] And Jaworska holds that the chief distinction between valuing and mere desiring is this:

> We think it would be a mistake to lose our current values—we hold our values to be correct, or at least correct for us. And this means that we can typically give a rationale for why we consider something valuable or good, usually by situating this value in a larger normative framework.

[32] Thus, I disagree with Jaworska when she claims that whatever a person cares about is "internal" to the person—representing (part of) her authentic or true perspective ("Caring and Internality"). Although I agree that concerns can possess agential authority without being reflectively endorsed, I do not think we should say that concerns have agential authority when they are consciously disavowed. For instance, I would not attribute agential authority to ingrained social mores that a person no longer endorses but cannot shake at an emotional level.

[33] Compare Tiberius, *Well-Being as Value Fulfillment,* 11–13; Dorsey, *A Theory of Prudence*, ch. 5.

[34] Dworkin, *Life's Dominion,* 206.

> Also, since values are the sorts of attitudes that we allow could be correct or incorrect, they are open to criticism and revision.[35]

I think this is basically the right approach.

The main disagreement between Dworkin and Jaworska, recall, is whether we should privilege a person's more global values. On this point, both writers have some piece of the truth. If we are asking which attitudes can possibly possess the sort of prudential authority associated with critical interests, then Jaworska is correct to say that a person's values need not be global. On the other hand, if we are thinking about a person who has a range of values, then his more considered values have more authority. And often our more global values *are* our more considered ones; they represent what we think is most important for us once we have tried to put everything into its proper perspective. This is not invariably true, however. A person might have certain idle notions about the good life, into which he has put nowhere near as much thought as his more immediate values. For this person, his superficial-but-global ideas about the good life should be attributed less prudential authority than the more local values that he takes more seriously.[36]

Some writers agree that valuing is central to both agency and defining a person's interests, but they deny that value judgment is central to the attitude. One might, for example, think that valuing is really a matter of having higher-order desires to maintain and act on certain more immediate desires.[37] I agree that our values often come into view when we reflect on whether we are really "behind" a particular feeling or desire. But the questions the agent is typically asking in such cases are normative or evaluative questions: "Is this really what I ought to be doing?" "Is this a good thing to care about?" "Are my feelings appropriate here?" and so on. Indeed, if these are *not* normative questions, it is hard to see why the mere fact that higher-order desires involve reflections on

[35] Jaworska, "Respecting the Margins of Agency," 115.

[36] Even if my values possess agential authority for me, they cannot all bear on my well-being (unless I am exceptionally self-centered). Most of us also value things for their own sakes and for other people. So how do we distinguish the welfare-relevant values? This is a version of the "scope problem" that also affects desire-satisfaction theories. Trying to resolve this issue is too large a task to undertake here. But, to mention a few possible responses, one might think that my welfare-relevant attitudes are those that (a) refer to my own life or necessarily presuppose my own existence (Parfit, *Reasons and Persons*, 494–95; Overvold, "Self-Interest and Getting What You Want"), (b) that I conceive of as involving my own good (Dorsey, *A Theory of Prudence*, ch. 5), (c) that are "warm" and "affective" (Heathwood, "Which Desires Are Relevant to Well-Being?"), or (d) that refer to the goods that are relevant to well-being on the correct objective theory.

[37] Frankfurt, "Freedom of the Will and the Concept of a Person."

other desires should give them any special authority.[38] This suggests that the perspective of the agent is (at least typically) defined or informed by his view about what, in his circumstances, is good or what he has most reason to do.[39]

Perhaps the most common objection to the cognitivist approach to valuing is that our judgments of value underdetermine what we personally value. As Samuel Scheffler says, "There are . . . many activities that I regard as valuable but which I myself do not value, including, say, folk dancing, bird-watching, and studying Bulgarian history."[40] But the kind of value judgment that is most relevant to the cognitive approach is not a belief about what might be worthwhile for someone, but what is worthwhile for me, given my particular character and circumstances. This kind of judgment will naturally take into account my particular abilities, my tastes, and my existing emotional attachments. But none of these facts about me constitute what I value since I might repudiate any of them or at least downgrade their importance. What I value—in the present sense—is what I think is good, fitting, or worthwhile in my circumstances, taking all of this into consideration.[41]

Might there not still be cases where the reasons are inconclusive? Perhaps I have two very different courses of life open to me, and I come to identify with one even though I have as much reason to favor the other. Or maybe the question is simply how much weight I ought to attribute to each priority. Do such cases not suggest there is a notion of valuing or identification that goes beyond the reasons we have?[42] This is an important theoretical question, but we need not settle it for our purposes. First, even if we grant that valuing can outstrip our value judgments in cases of rational underdetermination, it remains the case that our judgments about what is good or worthwhile play the central role in informing our values. Generally, a person's values do align with her view about what is best or fitting for her life. Second, even if reason is sometimes inconclusive about which values we ought to initially adopt, it is also the case that once our lives have been shaped by one set of values, this changes our circumstances and, thus, also the facts about what is really best for us.[43] For instance, even if there were initially no conclusive reasons for me to personally value philosophy more than history, the fact that my life has been shaped by valuing philosophy

[38] Watson, "Free Agency."

[39] Wallace, "Caring, Reflexivity, and the Structure of Volition,"; see also Bratman, "A Desire of One's Own."

[40] Scheffler, "Valuing," 21.

[41] Wallace, "Caring, Reflexivity, and the Structure of Volition."

[42] Raz, "Incommensurability and Agency"; Bratman, "A Desire of One's Own."

[43] Raz, *The Morality of Freedom*, ch. 14.

for many years generates new reasons to continue valuing it. What is important for the argument going forward is simply that a person is typically unwarranted in revising his values—particularly long-held values regarding central elements of his life—without good reason.

3. The Revision Model II: The Authority of Past Values

I have argued that, among the subjective attitudes possessing prudential authority, a person's values are the most important. Further, I have claimed that valuing involves beliefs about what is good, fitting, or worthwhile for one's life and that changing one's mind in a justifiable way typically involves some kind of reconsideration or responsiveness to those reasons. Let us now consider how the prudential authority of values works over time and how this relates to cases involving dementia.

3.1. Prudential Authority over Time

The puzzle about the relevance of past attitudes is commonly discussed in connection to desire-satisfaction theories. It might seem that, on such views, the essence of prudence is to accord as much weight to the fulfillment of future desires as present ones. But if desires matter equally whenever they are held, should I not also try to fulfill my past desires?[44] For instance, if we imagine a six-year-old who decided he would like to go on a roller-coaster on his fiftieth birthday, would the person, now fifty, have prudential reason to get on the ride, even though he really does not want to do so now?[45] If past desires continued to matter, then just as we often make present sacrifices for the sake of what we will desire in the future, we would have to accustom ourselves to making sacrifices now for what we used to want. We would also have more reason to fulfill desires we had entertained for longer. This means it would be imprudent not to keep track of how long you had wanted something. Some find these implications very counterintuitive and infer that the desire-satisfaction theory of well-being is untenable. True, some of the problems dissipate once we remember that desires are usually conditional on certain factual assumptions (which may change), or on their own persistence, or on getting some enjoyment from their fulfillment.[46] But we can at least imagine cases that do not involve those conditions. For instance, Derek Parfit tells us that, when young, he strongly wanted to be a poet one day. This desire, he assures us, was not conditional in

[44] Brandt, *The Good and the Right*, 249–50; Parfit, *Reasons and Persons,* 150–53.

[45] Brandt, *The Good and the Right*, 249.

[46] Parfit, *Reasons and Persons,* 151; Bykvist, "The Moral Relevance of Past Preferences."

any of the above ways. "Does my past desire," he asks incredulously, "give me a reason to try to write poems now, though I now have no desire to do so?"[47]

But what happens when we shift our focus away from mere desires, to a person's values? As Parfit observes, when desires involve value judgments, things look different from the first-person perspective. If your fundamental values change, it seems you must take yourself to have an improved view of the reasons that apply to you. In other words, you are committed to believing that your current values are more justified than your former ones. So, just as you do not think your former, now discarded beliefs have some enduring validity, likewise your former values seem to have no prudential weight from your current perspective.[48]

Things are different, however, from the observer's point of view. Since the observer does not necessarily share either the agent's past or present perspectives, he is not committed to supposing that the agent's present values have more warrant than the old ones. So, why should he recognize the agent's present values as possessing special prudential authority? The answer, I suggest, is this. We have already seen that a person's values have higher authority in defining what a person's good consists in. But if we are going to take a person's values seriously *as values*—as beliefs about what is best for her—we have to accept her current perspective as a revision of her former one. If we simply toss her former views on the scales with her current ones, we will not be treating these as genuine values at all. Without the possibility of value revision, the normativity of valuing is undercut. But, as I argued above, it is that normativity that lends values their special authority to speak for the agent.

You may object to this reasoning that, just as a person's current values involve a commitment to rejecting her past values, a person's past values involve a like commitment to rejecting her incompatible future values. Why should the observer treat the agent's current retrospective rejection of her past values any differently than the agent's former prospective rejection of her current values?

Well, you might think that the observer's default assumption should be that an agent's later values are better supported by reasons than her earlier values

[47] Parfit, *Reasons and Persons*, 157. Some writers say that satisfying past desires *does* promote well-being (e.g., Baber, "Ex Ante and Post Hoc Satisfaction"; Sarch, "Desire Satisfactionism and Time"; Dorsey, *A Theory of Prudence*, ch. 8). But suppose a person, now permanently unconscious, long wanted his life extended no matter what, but had a change of heart last year. These views surprisingly say that (all else equal) the older, longer-held desires count for more than his more recent desires.

[48] What about values I predict I will hold in the future? If I expect I will later be better situated to make an accurate value judgment than I am now, I should bring myself to value that thing presently. If I expect my future judgment to be worse, I should disregard it.

because she has had more opportunity to gather evidence and experience. This may be true, but it seems like this presumption will be undermined whenever there is independent reason to believe that the person's judgment has worsened. If this were the only reason to favor a person's present perspective, then bystanders *should* sometimes ignore a person's present values in favor of her former, more reasonable ones. This would not be so very surprising if we were just talking about temporary judgment shifts due to weakness of will. But, here, our focus is on stable changes in a competent agent's values.

A better explanation of why a person's present values should count from the third-person perspective concerns the temporal nature of agency. We care about a person's values because we think it is important to know how a person has made up her mind. But making up one's mind is something done in time, taking off from previously held conclusions or unreflective attitudes. My claim is that you just cannot pay due regard to what a person *does* in making up her mind about her values unless you see her later judgments as supplanting her earlier ones. To look at values in a tenseless way obscures the person's active role in forming and revising them.

3.2. Answerability for Value Change

Appreciating why we should generally recognize a person's present values as authoritative throws into relief the exceptional cases in which it makes sense to continue to recognize the authority of a person's former values. Usually, we treat a person's current values as revisions of her past values. When ordinary agents without significant impairments discard values, the fact that they no longer hold the same views is something that they are directly answerable for since they can sufficiently understand and appreciate the considerations on both sides of the matter. This "answerability" is a species of responsibility: it makes sense to ask a person who is answerable why she changed her mind, and this makes her eligible for judgments as to whether she exercised her judgment well or poorly.[49] A person can be directly answerable for his change of values, notice, even if he did not discard his old convictions in a deliberate way. Preoccupied with other things, a person may simply let his former ideals and commitments fade. But this person is nonetheless answerable for the change in view if he remained capable of rational reflection on the views he was drifting away from; it would still make sense to ask him why his mind has changed. In that case, we should still treat his current perspective as authoritative.

[49] My thinking about answerability is influenced by David Shoemaker, *Responsibility from the Margins*, though I am adapting the idea for my own purposes. I discuss the same idea under the heading of the "Answerability Approach" in my "Binding the Self."

By contrast, when someone loses the psychological capacity to understand and appreciate why she held her former values, she is not directly answerable for the change of view. The change in what she cares about is not up to her, even if she remains capable of valuing or caring about other things.[50] And so, the reason we usually have for disregarding a person's past values—that this is part of recognizing the authority of judgment—no longer holds. We cannot sensibly regard her current outlook as a revision of her past one. That is why, contrary to Jaworska's suggestion, losing one's values as a direct result of impairment is not comparable to changing one's mind. As Dennis McKerlie observes, the dementia patient in this circumstance "did not change her mind, the disease changed her mind. It stripped her of the ability to even understand her old point of view."[51] Therefore, we have reason to think of those former values as still valid for her, still legitimately her own.[52] They are, as I will say, "enduring interests."

My central claim, then, is that people living with moderate-to-severe dementia sometimes have these enduring interests, even when they remain capable of valuing or caring about other things. Crucially, just because a person has enduring interests does not imply that we should ignore their current values or concerns. Whereas the Status Model asks us to think about prudential authority as belonging in a holistic manner to either a person's former or current perspective taken in its entirety, I am urging a piecemeal approach, according to which we identify the individual attitudes that can still properly be imputed to the person as authentically her own.

[50] Bearing responsibility for the event or process that changed one's view is not the key issue. If a person suffers a serious cognitive impairment due to a reckless accident, the person bears some responsibility for the injury. But if the impairment changes his values, he is not answerable for discarding his old convictions since it would not make sense to ask him to explain why he has changed his mind. See also the example of "Hollywood amnesia" below.

[51] McKerlie, *Justice between the Young and the Old*, 186. I am trying to offer a fuller explanation as to why this difference matters and make the case that recognizing enduring interests does not preclude attending to a person's present evaluative perspective.

[52] In "The Moral Relevance of Past Preferences," Krister Bykvist reasons that, just as it would be "unfair" to count *A*'s preferences involving *B*'s life unless *A*'s preferences are in harmony with *B*'s, so too it would be unfair to count the desire of A-at-t_1 for A-at-t_2 if it conflicts with the present-oriented desires of A-at-t_2. This avoids present-for-past sacrifices but counts the former desires of the now deceased or comatose. (I am not sure if he thinks present desires exclude past ones only if the person remains competent.) Bykvist and I reach some similar conclusions, but he thinks desires we held for longer matter more. This has the counterintuitive implication I mentioned in note 47 above. More generally, Bykvist analogizes our problem to interpersonal cases, whereas I start with the first-person perspective of the agent deliberating about whether he should revise his former values.

3.3. *Further Remarks*

Let me now clarify and expand on a couple of points. First, just what does it mean to say that a person must be able to "understand and appreciate" his former values? The clearest cases will involve losing the ability to grasp the concepts involved in one's former values. Jaworska gives the example of a Mr. O'Connor, who had been "a deeply religious man for whom thoughts of taking his own life or of withholding lifesaving measures for whatever reason were completely unacceptable." However, when his Alzheimer's disease undermined his "capacity for complex reasoning, most of his religious beliefs gradually faded away." Now, grieving for the loss of his wife, he says he is ready to die.[53] If Mr. O'Connor really cannot understand his former religious convictions anymore, then he cannot be responsible for the fact that he no longer holds them. In that case, we should give them some weight in thinking about his best interests.[54] The same would be true if the person retained some understanding of the value but had lost the ability to apply it to his own life.[55] Furthermore, deliberating about what is important to us also draws on our emotional capacities. If a person only has a very superficial grasp of his former values—somewhat in the way that a child might faintly grasp some of the things that adults deeply care about—then again, the person lacks the capacities for taking responsibility for a change in view.[56] For example, perhaps Mr. Amato still retains some command over the basic concept of privacy, but his dementia has affected his ability to feel any of his former concern about it.

It is true that a person without what we would ordinarily regard as an impairment may also say that he cannot understand his former perspective anymore. Suppose that Don formerly wanted to remain a bachelor his entire life, but having found himself with a wife and a child (not in that order), he is now profoundly satisfied with his current life and regards his former perspective as quite alien. Still, I think cases like this are distinguishable from those which undermine responsibility for a change in values. Even if Don no longer

[53] Jaworska, "Respecting the Margins of Agency," 107.

[54] This does not necessarily mean that his former convictions should have the final word. See below for my discussion of conflicts between a person's former perspective and current one.

[55] Similarly, a person could retain his convictions in a rigid way while losing his ability to reflect on their grounds. This, too, could deprive the person of the ability to adapt and revise his beliefs in light of new experience. In this circumstance, it may be appropriate to attribute less weight to the person's present convictions if it seems that the person‘s unimpaired former self would have applied the values in a different or more nuanced way.

[56] Understanding and emotional appreciation are also frequently regarded as requirements for decision-making competence (see Buchanan and Brock, *Deciding for Others*, 23–25).

feels the pull of his former values, he can comprehend what it was about living without familial responsibilities that once appealed to him, and he can recognize the relative significance someone (like his former self) might attribute to living that way. I grant that the difference between Don and someone like Mr. Amato may be a matter of degree, but that is generally true when we talk about the capacities that underlie responsibility. That there will be some hard cases (which may call for caution) does not show that we cannot often differentiate factors that do, and do not, undermine answerability.

Perhaps I should also say that I do not think it is essential to answerability that someone personally remembers having held certain values in the past. Suppose a person suffers from the sort of "Hollywood amnesia" where she remembers nothing about her past prior to an accident but is otherwise unimpaired. On the present account, this person's former values would *not* bear on her current interests so long as she remains capable of reexamining the sorts of reasons that grounded her former values (particularly if someone informed her of what used to be important to her). True, she was not responsible for her initial change of view, but she is answerable for the values she adopts henceforth.

Someone might accept that a person is harmed when he loses the ability to value what used to be important to him but think that, nonetheless, the interest disappears when the value does.[57] That is a possible view, but I am more persuaded by the intuitions on the other side. When we think about what is important for making our own lives go well, we probably suppose that some of these things could still matter even if we should lose the ability to care about them. Certainly, we would be willing to make sacrifices now to improve the prospects of fulfilling our values in such cases. Likewise, it does not seem extravagant to believe that one could set out to harm someone with dementia by undermining a project he had previously invested himself in but could no longer remember or appreciate—for instance, by destroying the manuscript of his last book.[58]

Given this last example, it may seem that the issue of enduring interests is no different from that of posthumous interests. That is not quite true, though there is some overlap. Some find it incredible that anything could affect a person's well-being if he no longer exists. But there is no "missing subject" in our dementia cases. Moreover, not only is the subject very much alive, but the subject is still the kind of being who can have interests grounded in his current perspective. For this reason, dementia cases can involve apparent conflicts

[57] Compare Luper, "Posthumous Harm."

[58] Someone might accept that destroying the manuscript is contrary to the individual's interests but deny that this affects her *well-being*. I will not quibble about the word. If you think that there are non-welfare interests, then I am happy to say that I am talking about those.

between the earlier and later perspectives, which again makes them unlike posthumous-interest cases.

4. THE REVISION MODEL III: THE IMPORTANCE OF WHAT WE USED TO CARE ABOUT

According to the Revision Model, a person's former values can still impinge on what is good for her, but only if she is not answerable for having discarded those values. How much, though, do these former values really matter in practice?

First, it is evident that people's former values, properly interpreted, do not always apply to their later lives. Some of the things we value are valued because of the way they make us feel. If those activities no longer have the same effect, then the former values will no longer be relevant.[59] There are many activities, similarly, that we only value doing on condition that we will continue to competently perform them. Again, a person might want her life to contain certain goods (like having a career or children) without feeling they must be present in every part of her life. And some values might be specific to particular life stages: Jane might have hated being highly dependent on others in her sixties but accept this as tolerable in her nineties. In other cases, a person has quite definite views about what she wants for her later life, but these are based on predictions or expectations that are not born out. People may, for instance, be overly pessimistic—or overly optimistic—about what life with dementia will be like.[60] Value judgments based on incorrect predictions should be accordingly discounted.

If, moreover, we think that well-being is mainly a matter of being receptive to what is objectively valuable, we will want to be convinced that what a person valued really is of some independent worth. Suppose Friedrich despised the elderly and thought that, when he himself reached old age, he would have no worth and his simple concerns should be disregarded. We might find that value judgment so unreasonable that we discount it or even entirely set it aside.

[59] Wendy Mitchell, living with early-stage dementia, writes of her former self: "We would not get on now, you and I. . . . We like different things. You live the work and bustle of a busy city, whereas some days I lose hours just looking out of a window at the view" (*Somebody I Used to Know*, 9). But, even if Mitchell could not understand and appreciate her former value for city dwelling, that former value would not be relevant anymore (on the present account) if it was connected to the way the city used to make her feel. (I owe this reference to Walsh, "Cognitive Transformation, Dementia, and the Moral Weight of Advance Directives," 60.)

[60] As Dresser observes in "Dworkin on Dementia."

When *are* a dementia patient's former values likely to be of lasting relevance? One category involves a person's goals or projects. Now, generally, a person must accomplish her own goals if they are to count as achievements. But, since accomplishments can be vulnerable to being undermined, we can have an enduring interest in their protection. In other cases, a goal has been *all but* accomplished, and it is only necessary for someone else to take the final step or fulfill her part so that past labors are not in vain.[61] A second category involves the maintenance of certain personal connections. Someone may, for instance, think it is important that grandchildren continue to visit, even if one day they seem like strangers to him. A third category involves cases in which a person sees something as important to a larger group, community, or tradition which she identifies with. Thus, Ms. Bell saw herself as a part of a family and thought it was important for certain heirlooms, which represented that family, to be passed down from one generation to the next. A fourth category relates to certain ways in which a person wants to be treated. While it is true that a person might cease to care about things like privacy, dignity, and bodily integrity, it may well have been previously important to the person that these things continue to be respected throughout her life. A fifth category relates to personal ideals. For these to be of lasting importance, they cannot be like ordinary moral standards, which do not seem to apply to a person to the extent that she is no longer accountable for her conduct. But not all ideals are like this. For example, someone might think it is important to remain faithful to his wedding vows, even if he should one day forget that he is married. Or, again, Mr. Amato might have formerly thought it important that others keep him clothed in public, even if he should cease caring about this, since walking around disheveled and half naked is not consistent with his conception of dignity. Of course, I acknowledge that some people would not want us to make decisions for them when they have impairments in the future in line with what they currently care about. "Just let me be who I am at that time," they may insist. But this value might itself be understood to be an enduring interest that should guide our interactions with them in the future.

As I have said, just because a dementia patient's former values have enduring relevance for her current interests, it does not follow that we should ignore her current concerns. These can still anchor critical interests. This is a key difference between the Revision Model and the Status Model. A person who cannot do many of the things he once enjoyed might discover simpler goods he had previously neglected. Suppose dementia deprives a man of his capacity to care about his past scholarly projects, but he comes to develop a deeper appreciation

[61] Compare Portmore, "Welfare, Achievement, and Self-Sacrifice."

for simply being with loved ones.[62] Acting in this man's best interests might involve both seeking to bring his projects to completion and ensuring he has opportunities to be with his family.

That is an easy case, in a way, since it does not force any trade-offs between a person's former values and what matters to him now. Part of the appeal of the Status Model is that it avoids those trade-offs altogether. But it does so at the cost of simply ignoring values and concerns rooted in either the person's past or present perspective. Moreover, it is not as if the Status Model can entirely avoid making other trade-offs. For one thing, the decision-maker will still face trade-offs between the interests rooted in a person's former values and his experiential interests.[63] Further, whether we are privileging the person's past or present perspective, there will still be trade-offs we have to make between different values internal to that perspective.

If the problem seems more intractable when trying to balance a person's past and present values, that may be due to the way that we deliberate when making decisions for others. In hard cases, at least with someone we know well, we do not identify the person's values individually and then assign them various weights. Rather, we imaginatively project ourselves into the other's place and ask ourselves how she would feel about things given her whole constellation of values. In essence, we run a mental simulation from the other person's perspective. But how can we reason from another's point of view if we must take into consideration values and concerns that are rooted in different perspectives?

We should not exaggerate the problem. A similar issue arises when we must make decisions for children with both immediate and long-term consequences, which requires balancing interests rooted in the child's current and future perspectives. But most people think that, however much uncertainty it may involve, this task of balancing the child's current and future interests can be done. Indeed, it would be strange to accept, in this context, something like the Status Model and say that we should not consider the values the child is likely to have in the future as long as she values some things now.

But I want to suggest that there *is* a perspective we can imaginatively construct that helps us reason about the dementia patient's interests. We begin with the person as she was just before the onset of her impairments. Then

[62] Christine Bryden, writing about her experience with dementia, observes: "As my cognitive ability fades, I have felt a greater sense of emotional connection within the community, and an increasing relationship with the divine" (*Will I Still Be Me?* 80).

[63] Dworkin apparently assumes that a person's critical interests are always more important than her experiential interests—but that is implausible. The smallest improvements in a person's critical interests do not matter more than (for instance) avoiding intense and prolonged pain. See Shiffrin, "Autonomy, Beneficence, and the Permanently Demented," 209.

we imagine the unimpaired person with us, observing her current situation, empathizing with her current priorities, and we imagine asking her what she would want for her life, given what she knows now, if she knew that she was about to take the place of her impaired self.[64] This may seem to grant a certain priority to the person's past perspective since we are asking how this perspective would change after encountering her present self. But it is hard to see what alternative there is since only the person without impairments can consider both perspectives. In any case, by adopting this approach, we recognize that there is at least one respect in which the person with the impairment is in a better epistemic situation since she is acquainted with what it is actually like to live with dementia—something her former self probably could not adequately appreciate. Ordinarily, when a person gains significant new experiences, that is an occasion for her to revisit and possibly revise her values. But, in this case, the new experience is precisely what prevents her from taking the necessary broader point of view. So, that is why we ask whether the person as she was before the impairment would have changed her mind about anything if she could encounter her current self and foresee taking her place.[65]

5. The Presentist Model

So far, I have been discussing positions that accept that a person's former attitudes can sometimes bear on his best interests. But what about the *Presentist Model*, which holds that only a person's present attitudes can ever affect his present well-being?[66] How compelling is this approach?

Well, you might accept the Presentist Model because you endorse a version of the "experience requirement": that nothing can affect a person's well-being at

[64] This echoes (but departs from) Railton, "Facts and Values."

[65] In "Cognitive Transformation, Dementia, and the Moral Weight of Advance Directives," Emily Walsh argues that we cannot simply assume that changes in a dementia patients' preferences are due to the degradation of their faculties; they may be due to having undergone the "transformative experience" of living with dementia. I agree that we must consider how the dementia patient's experiences might have changed her mind if she were capable of reflecting on both her former views and her current ones. I disagree with Walsh, however, when she suggests that we can never make a reasonable inference as to what caused a change in perspective. That pessimism seems especially implausible when we look beyond the medical context to cases in which a person has apparently ceased to care about loved ones or past projects. There is much more one could say about transformative experiences and dementia, but the issues are too intricate for me to take them up here.

[66] Of course, a person's future wants or values can matter insofar as they bear on her future well-being, but a person's past attitudes cannot matter because we cannot (on this view) affect our past well-being.

a time unless that thing is consciously experienced as good or bad by the person at that same time.[67] Such a view does not *identify* well-being with the quality of the experience (as hedonism does); it only regards that experience as a necessary condition on something's affecting a person's interests. [68] So, even if it is generally good for me to get what I value, it will not benefit me if I never find out that my value was fulfilled. Taking a symmetrical view about harms implies the truth of the old adage, "What you don't know can't hurt you" (at least, not directly).

Those who are skeptical of the experience requirement might appeal to a case from Thomas Nagel, which (following some later retellings) we can dub the "Deceived Businessman." Nagel says that if you think that what you do not know cannot affect your well-being, then "even if a man is betrayed by his friends, ridiculed behind his back, and despised by people who treat him kindly to his face, none of it can be counted as a misfortune for him."[69] If that seems like the wrong answer, then we should reject the experience requirement.

But that is not the only possible ground for the Presentist Model. One might instead invoke an idea I have also drawn upon: that a person cannot directly benefit from something unless it resonates with him in the right way (at some level). Crucially, this simple formulation of the "resonance requirement" leaves matters of timing vague. *When* must a person have the right pro-attitudes toward something if it is going to directly benefit him? Jennifer Hawkins, in an important article, defends a concurrentist interpretation: something cannot directly benefit a person at a given time unless he responds positively to it at that same time or *would* do so were he aware of it. (She accepts a symmetrical view for direct harms.)[70] The implication is that if a person does not care about something anymore, then that thing can no longer bear on her interests. However, since the principle only requires that a person would respond to something in a certain way if he were aware of it, this approach can allow that the Deceived Businessman has been made worse off. Still, Hawkins keeps the hypothetical on a short leash: we are not supposed to idealize the subject's responses in any way beyond imagining her aware of that which she is ignorant.

We can make this view more concrete by considering two versions of a story.

> *Forgiveness 1*: Jay, struggling with alcohol, abandoned his family twenty-five years ago. Feeling remorseful, though, he recently wrote his

[67] See Griffin, *Well-Being;* and Sumner, *Welfare, Happiness, and Ethics.*

[68] Thus, the experience requirement is not so vulnerable to the most widely accepted problem for hedonism: that most of us would not consider a pleasant life plugged into an "experience machine" to be a good one (Nozick, *Anarchy, State, and Utopia*, 42–45).

[69] Nagel, "Death," 4. See also Kagan, "Me and My Life."

[70] Hawkins, "Well-Being, Time, and Dementia." She dubs this the "non-alienness principle."

> daughter, Kate, a long apology. He was disappointed when he did not receive a response, but he felt that if only Kate would forgive him in her heart, he would have achieved something in making amends. When Kate first received the letter, it only made her angry again. But soon thereafter, she discovered that her own teenage son was dealing with a drug problem. This softened her feelings toward her father. But with all that was going on, she did not reach out to him right away. Then she received word that he had died.

No doubt it would have been better for Jay if he had learned that Kate forgave him. But what mattered most to him was his daughter's feelings toward him, not his own knowledge of those feelings. So, although the experience requirement would say that Kate's change of heart could not have been good for Jay, Hawkins would take the opposite view since he would have been pleased if he *had* known how she felt. Now consider a second version of the case:

> *Forgiveness 2*: The story begins the same way, but now Kate does call her father as soon as she has the change of heart. However, Jay is so ill at the time he lets the call go to voicemail. He hears Kate say that she forgives him, and he understands what she is saying, but he is feeling so wretched that now the call just annoys him. Before he can recover, he dies.[71]

In this version, since Jay is aware of Kate's feelings but is unmoved by them, the concurrent resonance requirement says that Jay does not benefit.

I think we should reject this interpretation of the resonance requirement. When we imagine how Jay would have responded in Forgiveness 1 to Kate's change of heart, I think we implicitly represent him as sober, not fatigued, not depressed, not distracted with pain, and so on. Why? Because we want to know what Jay's true opinion is, and we assume that those states distort one's authentic perspective. But if that is right, we should also think about Forgiveness 2 differently. If Jay felt so bad that he was not "fully himself" when he learned that Kate forgave him, we should consider instead how he would have responded if he were feeling better.

Someone might deny that we should appeal to Jay's "authentic" attitudes even in the first case; we should only consider how Jay would actually respond given the condition he was presently in. If Jay was ill when Kate had the change of heart, and he would not have been pleased if he had known about it, then we should say that her new feelings are not good for him *at that moment.* But, oddly, this implies that if Jay cycles through a series of moods, then his well-being will

[71] This resembles the Princess Lovely example in Hawkins, "Well-Being, Time, and Dementia," 529–30.

wax and wane depending on whether he would then welcome Kate's forgiveness. My sense is that if something can be intrinsically good for someone without it entering his awareness, then the more plausible view is that the welfare effect depends on his stabler underlying evaluative perspective, not his transient moods. In that case, it seems like it is also a person's authentic attitudes that matter when he *is* aware of the event.

We could, however, revise the concurrent resonance requirement so that it refers to how a person would respond to something if she knew about it *and if she were judging it in accordance with her authentic evaluative perspective*. If we assume that a person's authentic or authoritative perspective is grounded in her current psychological dispositions, it would still follow that her former values can have no bearing on her current interests. In my opinion, this is the most plausible basis for the Presentist Model.

Nevertheless, once we accept that it is a person's *authentic* perspective that matters, it is but a short step to the Revision Model, which says that a dementia patient's authentic standpoint may be grounded (in part) in the values she used to hold and never genuinely revised.[72] Moreover, I would suggest that the Revision Model is closer to everyday ways of thinking. If your mother, once devoted to her grandchildren, lost her ability to remember or recognize them, you probably would not declare, "She has obviously changed her mind about how important her grandchildren are to her." No, you would more likely say, "Sadly, she just is not herself in that respect anymore." That, anyway, is what I would want loved ones to say about *me* in that circumstance.

6. FINAL THOUGHTS

Let me retrace the outline of the Revision Model. When making decisions for a person without significant permanent impairments, we properly focus on his current values since these represent the outcome of his deliberations to this point about what is important in his life. But if, due to an impairment like dementia, a person cannot understand and appreciate the considerations underlying some of his former values, then the fact that he no longer holds these is not really up to him. Since the person never really revised his values, we should presumptively regard those values as still a part of his last authoritative

[72] If well-being can be higher or lower at different times, then at what point in the person's life is his well-being promoted if his former value is realized now. Is it now or is it in the past? I would say that the person is better off once both the relevant attitude and the state-of-the-world have obtained. So, if a former value is fulfilled now, then the person benefits now. Likewise, if he comes to value something now that happened in the past, he also benefits now. See Lin, "Asymmetrism about Desire Satisfactionism and Time."

verdict about his own good. If those values directly bear on his current interests yet seem to conflict with his current values or concerns, we should ask ourselves what the formerly unimpaired person would have wanted for himself if he were here with us now, sympathizing with his current self and that self's distinct perspective, and anticipating taking the place of his current self. A virtue of this approach, I have said, is that it explains why a person's past values may be relevant in cases involving impairment but ordinarily are not.

Someone may object that this account is inconsistent with treating those living with dementia as persons worthy of respect and concern because it emphasizes what people with dementia are losing, rather than who they are now. As such, it might seem stigmatizing or depersonalizing. What people living with dementia want and need is not that we become the caretakers for their past values and projects but that we take their current perspectives seriously. Consider, for instance, the words of Christine Bryden, who has been living with dementia for many years: "I am who I am now, and meaning is what I can find in this present moment. My narrative results from finding meaning in life and developing a sense of identity in the present moment, not based on events in the past."[73] This is a serious concern. The move toward a "person-centered" approach has been the most notable trend in the ethics of caring for people living with dementia over the last couple of decades.[74] If the present account is inconsistent with that approach, that is probably a fatal strike against it.

So, let me respond. First, I hope it goes without saying that I have been assuming throughout that the individual with dementia remains *a person* in the sense of deserving our full moral consideration.[75] My question has simply been whether any of the person's interests might be rooted in her former evaluative perspective. This should be no more depersonalizing than asking whether a person has an interest in fulfilling his future values or desires. Further, I trust that I will not be interpreted as saying that anyone who is diagnosed with dementia (much less diagnosed with a disease that causes dementia) will be unable to authoritatively change their minds about what is important to them. Someone may have significant cognitive impairments, including difficulty remembering parts of their past and in articulating what matters to them now, and yet sufficiently retain the rational and emotional abilities that make

[73] Bryden, *Will I Still Be Me?* 99. I am grateful to an anonymous reviewer for recommending I address this issue and for directing me to Bryden's work.

[74] See Kitwood, *Dementia Reconsidered*; Sabat, *The Experience of Alzheimer's Disease*.

[75] Rejecting the Status Model does not imply that people living with dementia lack moral status. Thus, I demur from the way that Jaworska defines "full moral standing" as inconsistent with recognizing the current prudential value of a person's former values (in "Caring and Full Moral Standing").

them fully answerable for discarding their former values. In all but the most severe cases, to reasonably infer that a person has enduring interests will require having communicated with the individual about her past perspective. Simply observing that the person has certain impairments is not enough.

I also want to reiterate that my main aim has been to offer a theory about the interests that people have—not a theory about the conditions under which we ought (all things considered) to try to promote those interests. It seems to me that most of the concerns about stigmatization are concerned with how we treat people, or the ungrounded assumptions we make about them, not with the sorts of interests they have. Indeed, any act of justified paternalism must (at the minimum) do more good than harm. Just because a person has certain enduring interests, it does not follow that interventions to promote those interests are always justified, particularly if those might inflict significant harms on the person.

More to the point, I consider it a virtue of the Revision Model that, as compared to the Status Model, it does not justify the relevance of a person's former values on the grounds that the patient can no longer act from a "sense of self." On the contrary, I have tried to stress the ways that the current perspective of the person with enduring interests also continues to matter on my approach. Therefore, the Revision Model is very far from regarding the person living with dementia as an "empty shell" of her former self.[76]

In fact, I think that attending to a person's former values *is* often a good way of showing concern and respect for the present person. Consider, first, the individual who is in the process of losing her memory of the values that long structured her life but who remains capable of recalling them with the aid from others. I think it quite plausible that helping this person retain a connection to her past concerns is intrinsically good for her. But why would we think that if we do not suppose that the person retains an interest in values she never truly discards? Why not just think she is in the midst of changing her mind and go with the flow?

By the same token, I think that many people in the early stages of dementia who are anxious about losing important parts of their self-conception would be comforted in knowing that, if they should eventually lose a grip on some of these core values, their loved ones will remember, honor, and (if necessary) attend to them in their stead. In this way, a person's biographical identity can be upheld through relations with others, even while one is losing the ability to preserve it oneself.

So, I do not think that the approach defended here is guilty of depersonalizing or stigmatizing people living with dementia. Of course, I cannot *refute* the

[76] A concern voiced by Bryden, *Will I Still Be Me?*

view that respect for personhood demands that we attend solely to the present perspective of the person before us. That might be a fundamental intuition for some. My own approach has been guided by an attempt to combine two ideas. The first is that, because dementia is something that happens to people who have built their lives around certain goods, the interests rooted in those goods can endure even when the person loses the ability to appreciate them. The second is that, even when people do have these enduring interests, that does not mean that their current perspective on what matters to them is unimportant. Whether the Revision Model is compelling is, of course, up to the reader to decide. But I think it deserves a hearing, as it is rooted in some common—and I would venture quite humane—ways of thinking about the interests of people living with advanced dementia.[77]

University of Toronto
andrew.franklin.hall@utoronto.ca

REFERENCES

Baber, H. E. "Ex Ante and Post Hoc Satisfaction." In *Time and Identity*, edited by Joseph Keim Campbell, Michael O'Rourke, and Harry S. Silverstein, 249–67. Cambridge, MA: MIT Press, 2010.

Brandt, Richard B. *A Theory of the Good and the Right*. Oxford: Clarendon Press, 1979.

Bratman, Michael E. "A Desire of One's Own." *Journal of Philosophy* 100, no. 5 (2003): 221–42.

———. "Reflection, Planning, and Temporarily Extended Agency." *Philosophical Review* 109, no. 1 (January 2000): 35–61.

Brudney, Daniel, and John Lantos. "Agency and Authenticity: Which Value Grounds Patient Choice?" *Theoretical Medicine and Bioethics* 32, no. 4 (August 2011): 217–27.

Bryden, Christine. *Will I Still Be Me?* London: Jessica Kingsley Publishers, 2018.

[77] Work on this paper was supported by a postdoctoral fellowship funded by the Templeton Foundation ("Love and Human Agency") as well as by an Insight Grant from the Social Sciences and Humanities Research Council. I am grateful to much helpful feedback I have received over the years from audiences at Binghamton University, the University of Toronto, the Centre for Ethics at the University of Toronto, the New York University Ethics Forum, and the 2022 "Ethics, Agency, and Personhood in Dementia" conference in Copenhagen organized by the University of Southern Demark. I also want to individually thank Laura Franklin-Hall, Thomas Hurka, L. W. Sumner, Sergio Tenenbaum, Kevin Vallier, Daniel Viehoff, two anonymous referees for this journal, and especially Agnieszka Jaworska.

Buchanan, Allen E., and Dan W. Brock. *Deciding for Others: The Ethics of Surrogate Decision Making*. Cambridge: Cambridge University Press, 1989.

Bykvist, Krister. "The Moral Relevance of Past Preferences." In *Time and Ethics: Essays at the Intersection*, edited by Heather Dyke, 115–36. Dordrecht: Kluwer, 2003.

DeGrazia, David. *Human Identity and Bioethics.* Cambridge: Cambridge University Press, 2005.

Dorsey, Dale. *A Theory of Prudence.* Oxford: Oxford University Press, 2021.

Dresser, Rebecca. "Dworkin on Dementia: Elegant Theory, Questionable Policy." *Hastings Center Report* 25, no. 6 (November–December 1995): 32–38.

———. "Life, Death, and Incompetent Patients: Conceptual Infirmities and Hidden Values in the Law." *Arizona Law Review* 28, no. 3 (1986): 373–405.

Dworkin, Ronald. "Foundations of Liberal Equality." In *Equal Freedom: Selected Tanner Lectures on Human Values*, edited by Stephen Darwall, 190–306. Ann Arbor: University of Michigan Press, 1995. Originally published in *The Tanner Lectures on Human Values*. Salt Lake City: University of Utah Press, 1990.

———. *Life's Dominion: An Argument about Abortion, Euthanasia, and Individual Liberty.* New York: Knopf, 1993.

———. *Sovereign Virtue: The Theory and Practice of Equality.* Cambridge, MA: Harvard University Press, 2000.

Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68, no. 1 (January 1971): 5–20.

Franklin-Hall, Andrew. "Binding the Self: The Ethics of Ulysses Contracts." *Ethics* 34, no. 1 (October 2023): 57–88.

Griffin, James. *Well-Being: Its Meaning, Measurement, and Moral Importance.* Oxford: Clarendon Press, 1986.

Hawkins, Jennifer. "Well-Being, Time, and Dementia." *Ethics* 124, no. 3 (April 2014): 507–42.

Heathwood, Chris. "Which Desires Are Relevant to Well-Being?" *Noûs* 53, no. 3 (September 2019): 664–88.

Jaworska, Agnieszka. "Advance Directives and Substitute Decision-Making." *Stanford Encyclopedia of Philosophy* (Summer 2017). https://plato.stanford.edu/archives/sum2017/entries/advance-directives/.

———. "Caring and Full Moral Standing." *Ethics* 117, no. 3 (April 2007): 460–97.

———. "Caring and Internality." *Philosophy and Phenomenological Research* 74, no. 3 (May 2007): 529–68.

———. "Respecting the Margins of Agency: Alzheimer's Patients and the Capacity to Value." *Philosophy and Public Affairs* 28, no. 2 (Spring 1999):

105–38.

Kagan, Shelly. “Me and My Life.” *Proceedings of the Aristotelian Society* 94, no. 1 (June 1994): 309–24.

Kitwood, T. M. *Dementia Reconsidered: The Person Comes First*. Buckingham, UK: Open University Press, 1997.

Lin, Eden. “Asymmetrism about Desire Satisfactionism and Time.” In *Oxford Studies in Normative Ethics*, vol. 7, edited by Mark C. Timmons, 161–83. Oxford: Oxford University Press, 2017.

Luper, Steven. “Posthumous Harm.” *American Philosophical Quarterly* 41, no. 1 (January 2004): 63–72.

McKerlie, Dennis. *Justice between the Young and the Old*. New York: Oxford University Press, 2013.

McMahan, Jeff. *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford University Press, 2002.

Mitchell, Wendy. *Somebody I Used to Know: A Memoir*. London: Bloomsbury Publishing, 2018.

Nagel, Thomas. “Death.” In *Mortal Questions*, 1–10. Cambridge: Cambridge University Press, 1979.

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

Overvold, Mark Carl. “Self-Interest and Getting What You Want.” In *The Limits of Utilitarianism*, edited by Harlan B. Miller and William H. Williams, 186–94. Minneapolis: University of Minnesota Press, 1982.

Parfit, Derek. *Reasons and Persons*. Oxford: Clarendon Press, 1984.

Portmore, Douglas W. “Welfare, Achievement, and Self-Sacrifice.” *Journal of Ethics and Social Philosophy* 2, no. 2 (September 2007): 1–29.

Raibley, Jason. “Well-Being and the Priority of Values.” *Social Theory and Practice* 36, no. 4 (October 2010): 593–620.

Railton, Peter. “Facts and Values.” *Philosophical Topics* 14, no. 2 (Fall 1986): 5–31.

Raz, Joseph. “Incommensurability and Agency.” In *Engaging Reason: On the Theory of Value and Action*, 46–66. Oxford: Oxford University Press, 1999.

———. *The Morality of Freedom*. Oxford: Clarendon Press, 1986.

Sabat, Steven R. *The Experience of Alzheimer’s Disease: Life through a Tangled Veil*. Oxford: Blackwell, 2001.

Sarch, Alexander. “Desire Satisfactionism and Time.” *Utilitas* 25, no. 2 (June 2013): 221–45.

Scheffler, Samuel. “Valuing.” In *Equality and Tradition: Questions of Value in Moral and Political Theory*, 15–40. New York: Oxford University Press, 2010.

Shiffrin, Seana Valentine. “Autonomy, Beneficence, and the Permanently Demented.” In *Dworkin and His Critics: With Replies by Dworkin*, edited by Justine Burley, 193–217. Oxford: Blackwell Publishing, 2004.

Shoemaker, David. *Responsibility from the Margins*. Oxford: Oxford University Press, 2015.

Sumner, L. W. *Welfare, Happiness, and Ethics.* Oxford: Clarendon Press, 1996.

Tiberius, Valerie. *Well-Being as Value Fulfillment: How We Can Help Each Other to Live Well*. Oxford: Oxford University Press, 2018.

Wallace, R. Jay. "Caring, Reflexivity, and the Structure of Volition." In *Normativity and the Will: Selected Papers on Moral Psychology and Practical Reason*, 190–211. Oxford: Clarendon Press, 2006.

Walsh, Emily. "Cognitive Transformation, Dementia, and the Moral Weight of Advance Directives." *American Journal of Bioethics* 20, no. 8 (2020): 54–64.

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

Journal of Ethics and Social Philosophy
vol. 27, no. 1 · march 2024

https://doi.org/10.26556/jesp.v27i1.1953
© 2024 Author

YOU AREN'T REALLY BLACK, YOU AREN'T REALLY WHITE

RACIAL DENIALS AND EPISTEMIC INJUSTICE IN THE BLACK-WHITE MULTIRACIAL COMMUNITY

Erica Preston-Roedder

Multiracial persons, e.g., people with parents of multiple races, are a significant demographic group within the us. Nevertheless, philosophical work on race has largely, and problematically, elided this group: we have ignored their distinctive racial experiences, and we have failed to deeply engage with the philosophical issues raised by multiraciality. This essay begins to correct that neglect by seeking to understand one aspect of multiracial experience—specifically, racial denials. A racial denial occurs when a person's description of their racial identity (e.g., "I am Black") is challenged or called into doubt. While monoracial individuals can generally assert their race without being challenged (e.g., "I am Black," "I am White," "I am Asian"), multiracial individuals cannot always do so. Upon asserting "I am Black" or "I am White," a multiracial person may be met with the rejoinder, "You aren't *really* Black" or "You aren't *really* White."[1]

Through a consideration of racial denials, this essay aims to demonstrate that, in many cases, multiracial individuals face a hermeneutically unjust epistemic environment. This unjust epistemic environment is significant because it can undercut a person's ability to understand and communicate her racialized experiences. To make this argument, I will carefully tease apart how different kinds of racial denials operate. My focus will be on illuminating the *epistemic injustice*

[1] There are a number of sociological treatments of racial denials among multiracial persons. For classic treatments, see Root, "The Multiracial Experience"; Hall, "Please Choose One." For more recent discussion, see Song, "Who Counts as Multiracial?"; Townsend, Markus, and Bergsieker, "My Choice, Your Categories."

Although this paper focuses on racial denials directed at multiracial people, it is important to note that multiracial individuals are not the only ones to experience racial denials, e.g., monoracial individuals with ambiguous racial appearance may also face racial denials.

involved in these racial denials. That is, I will be focusing on ways that multiracial individuals are damaged in their capacity as communicators and self-knowers.[2] Moreover, by providing a careful description of how epistemic injustice operates within certain racial denials, I will draw out a number of larger implications for how we might understand race and epistemic injustice generally.

Before I begin, here are several preliminary notes. First, for reasons of scope, this essay will focus on multiracial individuals with Black and White ancestry. Careful sociological work has highlighted the distinctive experiences of different multiracial groups. For instance, Strmic-Pawl has argued that persons of Asian-White descent are "closer" to Whiteness, and they thus experience their mixedness quite differently from those of Black-White descent.[3] In a different vein, Rudy Guevarra Jr., has argued that the historical influence of Spanish colonialism has created deep affinities between Mexican and Filipino culture; because of this, persons of mixed Mexican-Filipino descent have generally been well-accepted by both their cultures.[4] In light of work like this, it seems judicious to begin an inquiry into multiracial experience by focusing our gaze on a specific subgroup—namely, persons with one Black parent and one White parent.[5] While I suspect that much of what I say here will generalize to other multiracial groups, this should not be assumed. For the rest of this paper, I will use the term "multiracial" or "multiracial individual" to refer *only* to members of this subgroup. I will occasionally use the longer term "Black-White multiracial individual" to remind the reader of this focus.

Second, I aim to largely eschew the thorny question: What is race? Let us allow that there are races but be agnostic (for the most part) about the details—biology, social construction, ancestry, etc. I will have some remarks later to make about the metaphysics of race. For now, however, we need only the observation that many monoracial people are able to unproblematically claim a race (e.g., "I am White," "I am Black," "I am Asian"), but that people of mixed ancestry sometimes face racial denials—that is, their racial self-descriptions are rejected.

Finally, a word on the significance of this project. Decades ago, Black feminists, such as bell hooks, convincingly argued that feminist theory needed to move people of color "from the margins to the center." In a similar way, there

[2] This language paraphrases that found in Fricker, *Epistemic Injustice*.

[3] Strmic-Pawl, *Multiracialism and Its Discontents.*

[4] Guevarra, *Becoming Mexipino.*

[5] To be fair, this definition is too narrow. For instance, a person may have a mixed parent or a Black grandparent. For the purposes of this paper, however, it will be helpful to have a clearly defined population for "multiracial." For more on the debate about how to define "multiracial," see Song, "Who Counts as Multiracial?"; Alba, *The Great Demographic Illusion*.

is both a theoretical and ethical imperative for philosophy of race to center the lives of multiracial people. The theoretical imperative arises because, as amply demonstrated by the last decades of feminist work, reflection on the lives of those at the margins has tremendous potential to enrich our understanding. That is, by examining a less-often scrutinized sector of life (i.e., women of color, multiracial experience), we can gain perspective and insight with respect to issues of broad philosophical significance. In this case, I will argue that analyzing racial denials can add nuance to our understanding of racial and epistemic injustice.

More importantly, there is an ethical imperative. In the case of feminism, it was necessary for White feminism to become more inclusive because, at bottom, the lives of women of color are just as interesting and important as those of White women—and therefore deserve equally substantive philosophical engagement. Similarly, the lives and experiences of multiracial persons deserve sustained attention. If this is right, then philosophy of race has an ethical imperative to reflect seriously upon the philosophical issues that arise within the experience of multiracial people. Further, I would argue that part of "centering" multiracial people is to devote philosophical energy and attention specifically to those phenomena that matter *within the lives of multiracial people*. The focus of this essay—racial denials—reflects this conviction. Specifically, autobiographical and fictional narratives of multiraciality commonly include accounts of racial denials, elegantly articulating the pain, confusion, and racial self-scrutiny they engender. If racial denials matter in the lives of multiracial people, and if multiracial people are to be centered in philosophy, then there is an ethical imperative to subject racial denials to sustained philosophical treatment.

1. RACIAL IDENTITIES AMONG BLACK-WHITE MULTIRACIAL INDIVIDUALS

Before making sense of racial denials per se, we must first understand the racial claims that multiracial individuals are apt to make. How do Black-White multiracial individuals racially identify? Existing literature suggests that contemporary Black-White multiracial individuals identify in a wide variety of ways. For instance, Davenport found that 25 percent of Black-White multiracial college freshmen identified as Black, 5 percent identified as White, and the remainder designated their race as "other" or as both "Black" and "White."[6]

To better understand such findings, it is helpful to move beyond statistical data and incorporate first-personal accounts from autobiography and

[6] Davenport, *Politics beyond Black and White*, 49. Davenport's methodology is more fully described later in the book (192).

sociology.[7] One of the most thorough and sensitive investigations of racial identity among Black-White multiracial individuals was conducted by Rockquemore and Brunsma, who offer the following taxonomy.[8]

> *Singular Black Identity*: Multiracial individuals with a singular Black identity conceive of their racial identity as solely Black. For instance, Aisha has a White mother and Black father, and she strongly identifies as Black.[9] Aisha relates a personal history in which she has been largely rejected by her White family, and she describes herself as "looking mostly black."[10] She currently attends a mostly White college, where most people assume she is Black, and she has experienced multiple racist incidents.
>
> *Singular White Identity*: Black-White multiracial individuals with a singular White identity conceive of their racial identity as solely White. While it is uncommon for a Black-White multiracial individual to identify solely as White, it is not unheard of; as noted earlier, roughly one out of twenty contemporary Black-White multiracial individuals identifies as White.[11]
>
> As an example of someone who identifies as White, consider Michelle, the daughter of a Black father and White mother. Michelle grew up in an upper-middle-class home and went to schools that were almost all White. Her friends have mostly been White. She acknowledges that she is "part African American," but she ultimately identifies solely as White.[12] Rockquemore and Brunsma offer the following characterization of her reasoning: "Her logic for determining her racial identification is that she *looks* white, she is *identified by others* as white, she was *raised* in a white community, she is *culturally* white, and therefore, she *is* white."[13] In another telling passage, they write, "Michelle so deeply

[7] In addition to the obvious descriptive richness of first-personal accounts, many scholars have emphasized the centrality of first-personal narratives in personal identity and ethics. See works such as Taylor, *Sources of the Self*; Alcoff, *Visible Identities*; Appiah, *The Ethics of Identity*; MacIntyre, *After Virtue*; Schechtman, *The Constitution of Selves*; Lindemann, *Holding and Letting Go*.

[8] Rockquemore and Brunsma, *Beyond Black*. While Rockquemore and Brunsma's book is more than twenty years old, their case studies are particularly vivid. Subsequent work has largely validated the analysis they offered. See Renn, *Mixed Race Students in College* and "Research on Biracial and Multiracial Identity Development."

[9] Rockquemore and Brunsma, *Beyond Black*, 39–40. All names are pseudonyms, as assigned by the researchers.

[10] Rockquemore and Brunsma, *Beyond Black*, 39.

[11] Davenport, *Politics beyond Black and White*, 49.

[12] Rockquemore and Brunsma, *Beyond Black*, 41.

[13] Rockquemore and Brunsma, *Beyond Black*, 41.

and clearly self-identifies as white that she describes the act of claiming a Black identity on her college admissions forms as 'passing for black.'"[14] Michelle's language of passing for Black is striking in that it underscores her sense that she is not, in fact, Black.

Border Identity: Black-White individuals with a border identity may identify using terms like "mixed" or "biracial." Those with a border identity see it as an identity that is neither White nor Black, but a distinct way to exist racially: living between two racial identities. Rockquemore and Brunsma's most detailed case study of a border identity is Anthony. Anthony's father left his family when he was quite young, and he was raised predominantly by his White mother and her family. Anthony attends a predominantly White high school in a rural community in Ohio. Among the non-White students at his school, roughly half have multiracial families. Anthony and his multiracial peers strongly identify as biracial. Indeed, Anthony describes himself by saying, "I'm *not* black, I'm biracial."[15]

While Anthony sees his biracial identity as an alternative to being Black, other multiracial individuals and theorists have interpreted the term "biracial" or "mixed" as potentially inclusive of other racial identities. For instance, Tina Fernandes Botts has characterized Black-White multiracials as both "black and mixed," and the filmmaker Lacey Schwartz has described "biracial" as being a subtype of Black.[16] In these cases, a person treats a biracial identity as compatible with a Black identity.

Protean Identity: Individuals with a protean identity see themselves as shifting between multiple identities, depending on the setting. For example, when Mike was asked about his racial identification, he replied, "Well shit, it depends on what day it is and where I'm goin'."[17] Mike, the son of a minister, is comfortable shuttling between his town's all-White and all-Black communities. As he moves between these groups, he adjusts his behavior; in doing so, he is not just performing, or playing at, being Black and being White. He sincerely understands himself as Black when he is with Black individuals, as White when he is with White individuals, and as biracial when he is with biracial individuals. Moreover, he feels that others validate his identity as Black, White, or

[14] Rockquemore and Brunsma, *Beyond Black*, 42.

[15] Rockquemore and Brunsma, *Beyond Black*, 43. Anthony's desire to be "*not* black" may be ethically problematic. For a discussion of the ethics of rejecting Blackness, see Sundstrom, "Being and Being Mixed Race" and *The Browning of America*.

[16] Botts, *Philosophy and the Mixed Race Experience*, 6; Schwartz, *Little White Lie.*

[17] Rockquemore and Brunsma, *Beyond Black*, 47.

> biracial across these contexts. This ability to be authentically at home in multiple racial identities is something he values about himself. Because he is authentically at home in various identities, and because he shifts between them, he thinks of himself as genuinely being Black, White, and biracial at different times.

In reading Rockquemore and Brunsma's case studies, it is tempting to challenge or reinterpret some of the claims made by interviewees. Mike, for instance, says that his racial identity "depends on what day it is, and where I'm goin'," but one might object that he cannot possibly mean this literally. Race, after all, simply is not the sort of thing that changes based on the day or setting. As another example, Michelle acknowledges that her parentage makes her "part African American"; given this, it may seem incoherent for her to characterize herself as White. That is, one might object that a person simply cannot have an African American parent and also be White. These kinds of responses are examples of *racial denials*.

In the remainder of this essay, I will examine the phenomenon of racial denials. I will argue that certain kinds of racial denials can be understood as products of unjust epistemic environments, although the exact form of the injustice varies according to the case. My analysis will begin by considering racial denials that call into question complex racial claims, such as those made by Mike (section 2), before turning to racial denials which target claims of being singularly White (section 3) or singularly Black (section 4).

2. Racial Denials: Monoraciality and Immutability

To understand racial denials that are directed at those with complex multiracial identities, it is helpful to first characterize two common assumptions about race: *monoraciality* and *immutability*. Monoraciality refers to the assumption that a person can only be one race; immutability refers to the assumption that a person's race cannot change.[18]

Monoraciality is challenged when a multiracial person asserts that she is of more than one race. For instance, as noted above, Botts describes Black-White multiracial people as both "mixed and Black," although she feels compelled to defend this, saying, "despite popular understandings of race in the United States, racial identity need not be an either/or proposition."[19] Botts's defensiveness is not misplaced—because of monoraciality, multiracial individuals who claim more than one racial identity face racial denials. For instance, Caroline

[18] Daniel, Kina, Dariotis, and Fojas, "Emerging Paradigms in Critical Mixed Race Studies," 12–14.

[19] Botts, *Philosophy and the Mixed Race Experience*, 6.

Ware, who identifies as Black and biracial, describes being asked, "Which side do you identify with most?" That is, her peers attempt to reformulate her self-described identity in terms that are consistent with monoraciality.[20] More generally, sociologists Johnston and Nadal have described how monoraciality undergirds a pattern of microaggressions encountered by multiracial individuals, including the demand that "You have to choose. You can't be both."[21]

Another way in which some multiracial individuals challenge prevailing notions of race is by claiming that they can move from one race to another—that is, rejecting the assumption that racial identity is immutable. Mike is an example of someone who experiences, and describes, his racial identity as fluid. Similarly, we might consider this excerpt from an interview with a young woman, Jane:

> It was always "mixed" when I was growing up. I think as I've gotten older, there's been a bigger focus on being black because of hearing awful things that happen in the black community and to black people and just identifying with that and being so struck by it and hurt by it.... It varies on the situation. Like when people say discriminatory things about black people, I identify more strongly with being a black woman. And then when there are comments about being mixed-race, I comment on my experience with that.[22]

Other research has confirmed this pattern: for some multiracial individuals, racial identity is situationally dependent.[23] Indeed, fluid conceptions of race are not uncommon; for example, a recent analysis found that mixed-race adolescents were four times more likely to *shift* their race than to identify consistently over time.[24]

Let us allow that multiracial people sometimes make claims that challenge monoraciality and/or immutability. These kinds of claims can lead to racial denials, where a person's self-ascribed racial identity is rejected ("You have to choose, you can't be both") or challenged ("Mike can't *really* mean that race is fluid"). How should we understand the phenomenon of racial denials?

One way to take up this question is to use the concept of *hermeneutical injustice*, as developed by Miranda Fricker.[25] On Fricker's view, individuals draw

[20] Williams and Ware, "A Tale of Two 'Halfs.'"

[21] Johnston and Nadal, "Multiracial Microaggressions," 133.

[22] Davenport, *Politics beyond Black and White*, 85.

[23] See Renn, "Research on Biracial and Multiracial Identity Development"; and Davenport, "The Fluidity of Racial Classifications"

[24] Hitlin, Brown, and Elder, "Racial Self-Categorization in Adolescence."

[25] Fricker, *Epistemic Injustice.*

on communal resources, such as shared concepts, to describe their own and others' experiences. Roughly, hermeneutical injustice occurs when a community's conceptual resources unfairly lack important concepts; because of this, a person's ability to understand or communicate some aspect of their experience is diminished.

To illustrate the notion of hermeneutical injustice, Fricker describes the development of the concept of sexual harassment. Before this concept was available, women who were victimized by sexually inappropriate behavior struggled to make sense of their experiences. For instance, one woman, Carmita Wood, had a supervisor who repeatedly jiggled his crotch, brushed her breasts, and at one point forcibly kissed her on the mouth. However, without the notion of sexual harassment, "Wood was at a loss to describe the hateful episode. She was ashamed and embarrassed."[26] Wood faces a gap, or lacuna, in the community's interpretive resources, and this gap makes it difficult for her to understand and communicate her experiences. The presence of a lacuna is a key characteristic of hermeneutical injustice: the collectively available resources do not include the concepts necessary to adequately understand and describe certain important aspects of people's lives.

The notion of hermeneutical injustice is important for our purposes because it calls attention to the way that gaps in communicative resources constrain our ability to communicate and to understand *ourselves*. When Wood was harassed, she had trouble articulating the experience to others; beyond this, she herself struggled to make sense of what was happening. For Fricker, limitations on conceptual resources impact our ability to communicate our experiences to others, as well as our self-understanding.

In a similar way, multiracial people often struggle to make sense of their racialized experiences and to communicate these experiences in ways that are intelligible to others. Consider, for instance, the words of Elliott Lewis in his autobiography. After facing a racial denial by a local business owner, he writes, "I didn't have the words . . . the intertwining of race, color and ancestry had rendered me speechless. I had no vocabulary to respond confidently or effectively to questions about my mixed and matched family."[27] Mariah Root, the noted multiracial activist, seems to be responding to a similar vocabulary failure when, in her well-known "Bill of Rights for Racially Mixed People," she writes, "I have the right to *create* a vocabulary to communicate about being multiracial."[28] It is necessary to create a vocabulary precisely because there are gaps in the

[26] Fricker, *Epistemic Injustice,* 150.

[27] Lewis is quoted in McKibbin, *Shades of Gray,* 66–67.

[28] Root, *The Multiracial Experience*, 7 (emphasis added).

existing available linguistic and conceptual resources. And Tina Grillo writes, "We have no stable conventions for describing multiracial persons, at least none that match what we perceive to be reality."[29]

The framework of hermeneutical injustice helps make sense of these examples. Put simply, our communal resources lack the concepts that these individuals need to describe themselves. The speakers have difficulty understanding and expressing their racial experiences because they face an emaciated vocabulary. Without shared communicative tools, listeners, in turn, have difficulty making sense of their claims. As Fricker writes, "Hermeneutical injustice most typically manifests in the speaker struggling to make herself intelligible in a testimonial exchange."[30] Indeed, even when a speaker uses language quite plainly, as Mike does, there may be little uptake from his hearers. Because of the conceptual gap in our shared resources, listeners may attempt to reinterpret Mike's racial claims in nonliteral ways (e.g., as a joke, bravado, or something similar). In short, listeners issue a racial denial. Racial denials, then, are a symptom of a deeper problem: speakers face a lacuna such that they lack adequate concepts to fully describe some important realm of their experience.

Fricker's account of hermeneutical injustice includes a second important component. Specifically, hermeneutical injustice arises when (a) there is a gap in conceptual resources *and* (b) the gap is an unjust one. For instance, in the case of sexual harassment, the reason the concept was not yet available can be explained by the fact that women were hermeneutically marginalized; that is, they were systematically denied the ability to shape and contribute to the interpretive resources in the culture.[31] For example, women did not hold leadership in major media outlets—positions from which they might be able to exert influence on shared interpretative resources. Not every lacuna is due to marginalization or some other form of injustice. In some cases, as Fricker writes, a gap may simply be "a poignant case of circumstantial epistemic bad luck."[32]

If we are to understand the lacuna around multiraciality as a case of hermeneutical *injustice*, we must therefore investigate why, exactly, communal resources lack nuanced concepts to describe multiracial experience. Here, I want to suggest we take seriously the possibility that multiracial persons, particularly those reporting complex racial identities, have been hermeneutically marginalized with respect to questions of race. Multiracial individuals have been reporting, for decades, that they have a wide variety of racial

[29] Grillo, "Anti-Essentialism and Intersectionality."

[30] Fricker, *Epistemic Injustice,* 159.

[31] Fricker, *Epistemic Injustice,* 155.

[32] Fricker, *Epistemic Injustice,* 152.

self-understandings and that their self-understandings are sometimes characterized by multiple and/or fluid identities. Indeed, those who have attended carefully to this space (e.g., the sociologists who have studied this population) have routinely noted this fact. Nevertheless, multiracial testimony to this effect has failed to have a significant impact on ordinary, communally shared conceptual resources around race.

One possible explanation for this is that multiracial people, particularly those who exhibit complex identities, are marginalized with respect to shaping communal resources around race: that is, multiracial persons are assigned a subordinate role in the communicative practices that develop, establish, change, and reinforce notions of race. In particular, because multiracial people do not have "regular," "normal," "pure," or "unambiguous" racial identities, a multiracial person is not seen as having the standing to speak authoritatively about race—including, in particular, the question of what it is to be Black. This status, that of being authoritative with respect to some realm of experience, or of having the standing to speak to it, is a kind of power—a kind of power that may be denied to multiracial people in virtue of the fact that they are multiracial. This, then, is the sense in which multiracial people may be marginalized with respect to crafting communal resources around race: because they lack a "normal" racial background, they are seen as less authoritative with respect to questions of race.

If this is right, then we need to grapple with the power dynamics of who controls communal resources with respect to racial concepts in order to fully understand why norms of monoraciality and immutability remain unchallenged. Insofar as multiracial people are assigned a subordinate status with respect to crafting communal resources around race, this will form part of the explanation for why our collective resources for understanding race continue to be gappy. Of course, marginalization will only form part of the complete explanation. In particular, a full accounting must also describe the role such lacunae play in maintaining White-supremacist norms, a topic that forms the focus of the next section of this paper. Nevertheless, the ongoing epistemic marginalization of multiracial people vis-à-vis questions of race should also be considered. Were we to take seriously the authority of multiracial people to speak on matters of race, instead of issuing racial denials, our communal resources might prove more labile.

In summary, I have proposed that our hermeneutical environment has gaps and that these gaps may be partly due to the epistemic marginalization of multiracial people around questions of race. This analysis is important in two respects. First, it helps make sense of certain characteristic experiences of multiracial persons. For a multiracial person, this gap in hermeneutical resources may hinder one's ability to communicate certain important aspects of one's life, leading to racial denials. In addition, it may make more difficult one's project of *self*-understanding. Second, if I am correct that multiracial persons have been unjustly marginalized in shaping our conceptual resources around race, this gives us reason to critically reexamine our concepts. Once we listen carefully to *all* voices about race, not just monoracial voices, this may lead us to a more sophisticated vocabulary around racial fluidity and racial multiplicity.

To conclude this section, I want to touch on several issues. First, it is important to situate Fricker's work against a Black feminist tradition. Decades before Fricker's work, Black feminists raised concerns about who circumscribes and controls the conceptual resources around race. [33] For instance, Angela P. Harris has argued that White feminists are too apt to project their own understanding of gender onto Black women in ways that flatten, obscure, and demean.[34] Similarly, Patricia Hill Collins has argued that Black women are subject to "*externally*-defined stereotypical images of Afro-American womanhood" and that there is power and benefit to *self*-definition.[35] For Collins, it is politically and ethically unjust for one group of people to control the conceptual resources used to describe the experiences of some other group of people; in particular, it is unjust for White persons to generate derogatory stereotypes to characterize the lives of Black women.

This tradition, then, encourages us to think about epistemic injustice as a space in which one group of people defines and controls the conceptual resources used to understand another group. On the picture I have sketched, monoracial persons tend to dominate and control the conceptual resources around race: monoracial people define what race is and how it works. Further, monoracial people deploy these resources to describe and understand the experiences of multiracial people. This is particularly vivid in cases of racial denials: monoracial persons explicitly refuse a multiracial person the opportunity for self-definition. Reflecting on Fricker's work within the larger context of Black feminist thought brings this aspect of multiraciality into sharper focus.[36]

At the same time, there are important differences. In particular, Collins paints a picture on which just and liberatory concepts are already available within the Black community. That is, for Collins, the Black community has empowering images of Black womanhood, but these concepts are ignored or overridden by White outsiders; different communities operate with different

[33] A number of authors have noted the debt that Fricker's work owes to Black feminists. See, for instance, Anderson, "Epistemic Injustice and the Philosophy of Race"; Pohlhaus, "Relational Knowing and Epistemic Injustice"; and Dotson, "A Cautionary Tale."

[34] Harris, "Race and Essentialism in Feminist Legal Theory."

[35] Collins, "Learning from the Outsider Within," s17 (emphasis added).

[36] My thanks to an anonymous reviewer who encouraged this direction of analysis.

conceptual resources, and one community imposes its concepts on another. In contrast, I have painted a picture on which even multiracial people *themselves* lack adequate vocabulary to understand complex racialized experience. In this respect, the epistemic position of multiracial people is similar to that of women before the concept of sexual harassment: even the community most impacted by the phenomenon lacks the vocabulary to fully articulate it, both to themselves and to others.[37]

Finally, I want to revisit the thorny question set aside earlier: What is race? Racial terms are notoriously complex and admit of many different interpretations. One might worry that the project I have undertaken here makes substantive assumptions about the nature of race, and that those assumptions need to be brought to light. For instance, consider biological and ancestral accounts of race.[38] If these accounts are correct, it is simply false that race is fluid: ancestry and biology are fixed, and therefore cannot vary based on context. If so, it would appear that Mike is making a straightforward factual error when says his race "depends on what time it is and where I'm going." In contrast, on social constructionist accounts, it is feasible to envision sociopolitical roles as fluid and context sensitive. [39] Does my discussion therefore rest on a nonbiological notion of race?

While this is a natural concern, it can be laid to rest; my discussion aims to be agnostic, and a proponent of biological races can, indeed, endorse my account. To see this, consider two arguments—both of which point back to gappy hermeneutical resources. First, suppose that ancestral/biological accounts are correct and that Mike *is* saying something false. For the purposes of this essay, the important point here is that *he does so precisely because he lacks a sufficient vocabulary to describe his experience*. Mike is not trying to deceive us; he is trying—earnestly and sincerely—to communicate something important about himself. He simply lacks the vocabulary necessary to do so without uttering a purported falsehood. More generally, the aim of this essay is not to vindicate the truth of Mike's claim but rather to diagnose the state of conceptual resources available to Mike.

Second, Mike's claim is literally *true,* even on a biological account, if we take him to be speaking about racial identity—roughly, a person's subjective sense of their race in their thought, emotions, actions, and self-understanding. That

[37] Dotson, "A Cautionary Tale," offers a rich discussion of this contrast between Fricker and Collins.

[38] For instance, see Hardimon, "The Ordinary Concept of Race"; Spencer, "A Radical Solution to the Race Problem"; Andreasen, "The Meaning of 'Race'"; and Kitcher, "Race, Ethnicity, Biology, Culture."

[39] Haslanger, for instance, makes this context sensitivity explicit in "Gender and Race."

is, we can expand our conceptual resources to differentiate between race and racial identity.[40] For most monoracial people, there is no need to differentiate between race and racial identity: one is Black *and* one thinks of oneself as Black; one is White and one thinks of oneself as White. On the other hand, multiracial people sometimes need conceptual resources that tease race and racial identity apart. Whatever Mike's race is in biological terms (if, indeed, race is biological), his subjective sense of his racial *identity* shifts from context to context. The fact that most speakers lack the conceptual vocabulary to make this distinction is indicative of an important, ongoing gap in our hermeneutical resources.[41] (It is also, incidentally, a reason to value philosophy, and philosophy of race in particular.)

As both these responses illustrate, our focus is on diagnosing hermeneutical gaps. Questions of "What is race?" are, therefore, somewhat orthogonal. Perhaps it will turn out that some of the claims made by multiracial people are false—but this is to be expected when a person is working with inadequate resources. As philosophers, we can do useful work by identifying and diagnosing these gaps, and considering ways we might expand our conceptual resources.

3. RACIAL DENIALS: DENYING WHITENESS

In this section, I turn to a different kind of racial denial. As noted earlier, roughly 5 percent of Black-White multiracial individuals identify as White. For instance, Michelle (profiled above) has a Black and a White parent but identifies as singularly White. A substantive sociological literature attests to the racial denials faced by White-identifying multiracial individuals. As one woman relates, "I was never fully allowed to identify as white or as Caucasian because when people saw me, that wasn't what they saw."[42] In another study, an interviewee states, "People look at me crazy if I say I'm white."[43]

It is generally well-recognized that the notion of White excludes individuals of Black-White parentage. American racial categories are governed by the one-drop rule, under which one "drop" of Black blood is sufficient to render a

[40] Appiah and Gutmann, "Race, Culture, Identity"; and Appiah, *The Ethics of Identity.*

[41] I have focused this discussion on biological/ancestral accounts of race because their tension with racial fluidity is obvious. However, even on a social constructionist position, a notion of racial identity is needed. For instance, a multiracial person who appears White may be afforded White privilege and, in that regard, inhabit the social position of Whiteness. Nevertheless, she may identify as Black or biracial. Haslanger makes precisely this point ("You Mixed?").

[42] Davenport, *Politics beyond Black and White,* 78.

[43] Khanna, "Ethnicity and Race as 'Symbolic.'"

person Black. As Naomi Zack has pointed out, this means that racial categories are applied asymmetrically: having a Black ancestor is sufficient to make a person Black, but having a White ancestor does not render a person White.[44] The one-drop rule can be traced back to early American slavery, and it became widely accepted under Jim Crow.[45] More generally, this asymmetry is often put forward as a paradigmatic example of hypodescent—that is, American racial practices typically assign individuals of mixed heritages to the "lower" racial denomination.[46]

Given this, White-identifying Black-White multiracial individuals anticipate and encounter racial denials. How should we understand these racial denials, and do they arise out of hermeneutical injustice? The framework of hermeneutical injustice directs our attention to two key factors: Does a person like Michelle (1) face a hermeneutical gap, where (2) this gap exists because of injustice?[47] Both factors are relevant here. First, as just noted, the term "White" excludes individuals with significant Black ancestry; that is, there is no available concept in the US that allows one to describe the experience of existing as White, and that can be used by those with significant Black ancestry. Second, the *reason* that this gap exists is one grounded in injustice: the concept of Whiteness has been shaped by racial oppression. In particular, the asymmetry of the one-drop rule has ensured that individuals of mixed parentage are denied the privileges of Whiteness. Most notoriously, in the antebellum South, a child of a White father and Black slave was commonly deemed a slave.[48] Thus, the situation exhibits both features of Fricker's hermeneutical injustice.

Charles Mills has considered related issues in his essay, "But What Are You *Really*?" Mills argues that, given that races are not natural kinds, the rules that assign a person to a racial group, particularly a person of mixed parentage, will be arbitrary and politically motivated. For instance, while the US operates using hypodescent, Mills points out that *hyper*descent is an in-principle possibility, i.e., assigning a child to the highest racial group. Indeed, in some parts of the world, this possibility was actualized: in the Dutch East Indies, social elevation was the

[44] Zack, *Race and Mixed Race.*

[45] Davis, *Who Is Black?*

[46] See, for instance, Harris, *Patterns of Race in the Americas*; Davis, *Who Is Black?*; Daniel, *More Than Black*; Khanna, *Biracial in America*; Jordan and Spickard, "Historical Origins of the One-Drop Racial Rule in the United States."

[47] This analysis extends Fricker's account somewhat. A strict reading of Fricker's *Epistemic Injustice* suggests that hermeneutical injustice arises *only* in cases of epistemic or hermeneutical marginalization. The injustice of the one-drop rule is, however, not limited to marginalization.

[48] Davis, *Who Is Black?*

norm, and Dutch Asian offspring were treated as Dutch.[49] In general, rules for racial categorization, particularly the racial categorization of people of mixed descent, are arbitrary social creations—where these creations are shaped by the political interests of the elite. In the US, the conceptual architecture of hypodescent has served to hoard material and social advantage within White families.

In the previous section, my focus was primarily on the individual: in cases of hermeneutical injustice, one's ability to communicate to others may be impaired, and one's self-understanding may be damaged. While these issues also arise for denials of Whiteness, Mills's work draws our attention beyond the individual, i.e., to the broader social function of racial denials. Racial denials, particularly denials of Whiteness, serve to police and enforce existing racial categories. Insofar as these categories function, as Mills argues, to uphold the interest of political elites, racial denials are one mechanism by which White supremacy is stabilized.

More specifically, racial denials uphold extant racial categories by rendering invisible the very phenomena which might otherwise pose a challenge. That is, in a world of enforced hypodescent, Black-White persons "disappear" into existing racial categories; there is no need to revise or revisit our racial vocabularies. In this way, we face a kind of feedback loop: practices of hypodescent ensure that multiracial people are accommodated within existing racial categories, and the fact that multiracial people fit into existing racial categories tends to stabilize and legitimate these very categories.[50] Put differently, our gappy hermeneutical environment obscures certain aspects of our racial reality, which thereby leaves us with the impression that this racial vocabulary is adequate.

Such a picture may seem to suggest that we have compelling reason to accept Michelle's claim of Whiteness. After all, to issue a racial denial would be to uphold problematic racial categories and, ultimately, problematic practices of racial privilege. However, existing scholarship suggests that there are also important reasons to be wary of extending Whiteness in this way.[51] For instance, one might worry that multiracial persons who identify as White are moved by

[49] Mills, *Blackness Visible*, 52.

[50] I am indebted to an anonymous reviewer for pressing this point and offering some of the specific language employed here.

[51] There is extant scholarship critiquing multiracial persons who identify as other than Black, although this literature typically focuses on those who characterize themselves as biracial/mixed instead of Black. Of course, these same concerns can be extended to those who identify as White. See, for instance, discussion in Sundstrom, "The Browning of America"; Davenport, "The Fluidity of Racial Classifications"; Zack, "Race and Mixed Race"; and Elam, *The Souls of Mixed Folk*. The phrase "escape hatch" was coined by Degler, *Neither Black nor White*.

internalized anti-Black sentiment or opportunism. In a different vein, some have argued that allowing multiracial people to identify in this way might serve as an "escape hatch" from Blackness, thereby undermining Black solidarity and Black political power. Finally, insofar as the concept of White inevitably invokes a hierarchy between Whiteness and other races, one might find it implausible to suggest that one can promote racial justice by expanding the scope of Whiteness.[52]

This essay cannot fully assess the question of whether, in the final analysis, the balance of reasons favors extending the concept of Whiteness to persons such as Michelle. I can, however, offer a clarification. A person like Michelle faces a hermeneutically unjust environment: she lives in a society with conceptual gaps rooted in injustice. This is true regardless of Michelle's particular motives for claiming Whiteness. However, in recognizing this, *one need not conclude that the only or best course of action is to thereby extend the extant concept of Whiteness to Michelle*. Perhaps this is the best course, but other responses are available. In particular, one might try to develop new, more liberatory concepts in order to make sense of experiences such as Michelle's. For instance, we might understand Michelle's experience of Whiteness using a cultural account of race, similar to that proposed by Chike Jeffers.[53] Jeffers argues that, while races do function hierarchically, racial groups can also function as cultural groups; we might use his account to suggest that Michelle identifies culturally as White through White-identified hobbies, music, and the like. The details and adequacy of Jeffers's account are not significant here. Rather, my point is that our response to a hermeneutical gap can be larger and more imaginative than simply granting Michelle admittance into an extant and problematic concept of Whiteness. When a person like Michelle offers a racial claim, we are not limited to merely issuing denials or offering uncritical acceptance. Instead, we can take her claim seriously, using it as an opening to critique our existing concepts of Whiteness and to create a richer vocabulary to capture the complexity of racial experience.

4. Racial Denials: Denying Blackness

A very different form of racial denial arises around denials of Blackness. Although roughly one out of four Black-White individuals identifies as singularly Black, multiracial individuals are not always accepted as Black.[54] For

[52] For instance, both Haslanger and McPherson have treated hierarchy as central to Whiteness. See Haslanger, *Resisting Reality*; McPherson, "Deflating 'Race.'"

[53] Jeffers, "Cultural Constructionism."

[54] Davenport, *Politics beyond Black and White*.

instance, Sarah Ratliff relates an encounter with her roommate, who had failed to invite Ratliff to an event held by the Congressional Black Caucus:

> "Girl, you wouldn't have fit in there. You do know that membership is only open to African-Americans, don't you?"
> "Yeah, I know that. Why wouldn't I fit in?" I asked.
> "Girl, I know you think you are Black, but for real? Girl, you ain't really Black!" She left the room laughing and talking to herself. "Girlfriend, light as she is, thinks she's Black!"[55]

Sarah is the child of a White father and a mother who identified as Black, although her mother was also of mixed heritage. While both Sarah's parents were American, the family spent time in Nigeria, which is where Sarah was born. After Sarah's parents divorced, Sarah was influenced by a (presumably Black) man she terms her "surrogate father," who was active in the Black Panther movement and fostered her sense of Black pride. As the exchange above suggests, Sarah has an ambiguous appearance, and she can sometimes be seen as White. Sarah describes her Black identity as rooted in a sense of Black pride and solidarity, although racial denials such as these have led her to question whether she can legitimately lay claim to a Black identity.

Does Sarah face hermeneutical injustice? And in what way does her situation differ from Michelle's? To begin, we should note a key difference between the two cases: while it is generally thought that one cannot have significant Black ancestry and be White, it is fairly common to accept that one can have significant White ancestry and still be Black (i.e., the one-drop rule). Sarah's roommate, therefore, is using the term Black in a way that is contested and arguably nonstandard. Whereas Michelle faced a lacuna in communal conceptual resources, here the difficulty is that the conceptual resources are contested and fractured. There are multiple conceptions of Blackness, and Sarah and her roommate are using different definitions. For instance, we might construe her roommate as suggesting that Blackness has a necessary condition: one can only be Black if one's phenotype leaves one vulnerable to racism. Sarah, in contrast, has a conception of Blackness that does not include this condition.

If this is so, Sarah does not face hermeneutical injustice because there is no lacuna. Under the one-drop rule, which is widely (if problematically) accepted, Sarah's Black ancestry is a sufficient condition to render her Black; like many other Black people in the US, she has relatively light skin, but her family tree includes significant Black heritage, and she lives her life as a Black person. Thus, there is no *gap* in the conceptual resources available for Sarah to understand

[55] Ratliff, *Being Biracial*, 34.

herself: she can describe herself as Black by utilizing this widely accepted notion of Blackness.

More generally, notice that a racial denial cannot create hermeneutical injustice. Hermeneutical injustice is a social phenomenon, a characteristic of the collective conceptual resources available to a person to understand herself. A racial denial, in contrast, is a local phenomenon, often taking place between just two individuals. Put concretely, a single individual (e.g., Sarah's roommate), utilizing one of several contested interpretations of a concept, cannot thereby create a gap in the community's conceptual resources. In this case, despite the roommate's racial denial, other understandings of Blackness still exist—the conceptual resources within Sarah's community have not changed.

While a racial denial cannot, on its own, create a hermeneutical gap, it can play an important, and related, role. In particular, while the roommate's racial denial cannot change the fact that multiple definitions of Blackness *exist* in the communal resources, the fact that these multiple definitions exist does not mean that Sarah, herself, can *access* all of these definitions. As a practical matter, it will be difficult for Sarah to understand herself as Black if the definition of Blackness she relies upon is not accepted by those close to her. Thus, while it is important to note that a racial denial by an individual does not change the communal resources that *exist*, it is equally important to note that racial denials—when issued in cases where there are multiple, contested conceptions—can function to ensure that certain racial conceptions are not *accessible* to an individual.

This gap, between what exists in conceptual resources and what a person can use to understand their lives, occurs in other contexts as well. For instance, a new parent of a disabled child might be aware of advocates who frame disability as a mere difference—as opposed to a difference that makes one worse off—but the parent might not yet find such a reconceptualization of disability personally compelling.[56] Or, to take a more mundane example, a well-meaning teacher might characterize a teen's difficult experience in class as "a learning opportunity." In contrast, the teen might characterize the same experience as "an embarrassing failure." While the teen is well aware of his teacher's conception, he simply does not see his experience in that way. He is capable of deploying the concept of a "learning opportunity," but it is not one that he is able to adopt into his worldview nor use to sincerely interpret his own experience. Similarly, Sarah is likely aware that there are many definitions of Blackness in the communal resources. Her difficulty is that racial denials from those close

[56] For a philosophical defense of the status of disability as mere difference, see Barnes, *The Minority Body*.

to her have made it difficult for her to adopt into her own worldview an understanding of Blackness that would include herself.

Given that racial denials can make certain conceptions less accessible to a person, racial denials can make self-knowledge more difficult and can, therefore, harm the person in their capacity as a self-knower. That is, a racial denial issued at a person like Sarah harms her in various obvious senses (it wounds her feelings, it prevents her from entering into solidarity with other Black persons, etc.). Beyond this, a racial denial harms Sarah as a self-knower; insofar as she is unable to access and use socially available conceptions of Blackness, Sarah will be hindered in her capacity to make sense of her own racialized experience.

Given that these kinds of racial denials can be harmful, is the harm inflicted unjustly? In the exchange above, the roommate's attitude toward Sarah is dismissive: she fails to take into account, or at least minimizes, Sarah's ancestry and experiences. In failing to take Sarah's autobiography into account, she does not give her her due and, thereby, treats Sarah unjustly. She also fails to take into account that Sarah has a claim upon Blackness that is widely recognized on other notions of "Black." In using an interpretation of Blackness that harms Sarah, and in failing to take seriously her autobiography and her use of another communally available definition of Blackness, she treats Sarah unjustly. In committing this injustice, she does not change the communal hermeneutical resources, but her dismissiveness does unfairly and culpably inflict harm on Sarah.

As a final note, these discussions of racial denials may bring to mind the case of Nkechi Amare Diallo, better known as Rachel Dolezal. Diallo is of White parentage, although she identifies as Black. Many individuals (including many Black individuals) deny Diallo's claim to be Black—that is, they issue the racial denial "You aren't *really* Black." Can this account of racial denials help us understand Diallo's case?

Racial denials directed toward Diallo are very different from the racial denial that Sarah encountered. Sarah's case involved multiple conceptions of Blackness. In contrast, Diallo was not laying claim to some preexisting notion of Blackness, but rather trying to extend the concept beyond the boundaries that are currently available in communal resources. In this, her efforts have more in common with Michelle, who tries to extend the boundaries of Whiteness to include herself; both individuals propose using terms in ways that are not accepted by widespread community standards.

In the case of Michelle, I suggested that the situation is one of hermeneutical injustice because the historical reason that White is not available to her is rooted in racist oppression: defining White in this way served a historical and social goal of preventing Black-White multiracial people from having access to the goods that White people enjoyed. In contrast, while the concept of Black is not available to Diallo, this is not because of racist oppression. That is, there is no known, significant American history in which persons of White descent have been denied the moniker "Black" as a way of limiting material and social privilege. Thus, while Diallo may face a gap in how she can self-identify, a gap she experiences as painful, this gap is not one of hermeneutical injustice since it is not due to unjust historical circumstances.[57]

In summary, racial denials targeting Blackness are not best understood as arising from hermeneutical injustice. Because of the one-drop rule, notions of Blackness are widely available that allow multiracial individuals to claim a Black identity. Nevertheless, these racial denials can damage a person in her capacity as a self-knower in a different, albeit related, way: by making certain conceptions of Blackness less *accessible* to the person, in the sense that one is less able to adopt such conceptions into one's own worldview. More generally, consideration of these kinds of racial denials demonstrates the importance of the accessibility, and not just existence, of conceptual resources for self-understanding.

5. Conclusion

I have argued that, in many cases, racial denials arise from underlying and unjust gaps in the hermeneutical environment; in these cases, racial denials are a symptom of an unjustly gappy conceptual vocabulary. In other cases, such as denials of Blackness, the hermeneutical environment is robust, although a racial denial may still be significant: it makes a certain conception of race less accessible to an individual. In general, it is fruitful to reflect on racial denials insofar as they call attention to the question of, not just what race is, but who has epistemic power and authority to control conceptual resources around race—both historically and in contemporary times.

In addition to understanding the phenomenon of racial denials, I aimed to demonstrate that centering multiraciality can yield broad insights regarding the philosophy of race and epistemic injustice. For instance, reflecting on multiraciality should lead us to conceptualize race, or at least racial identity, as multiple and fluid. With respect to epistemic injustice, I suggested that we distinguish between cases in which conceptual resources do not exist versus cases in which resources are inaccessible to the individual; this distinction is also useful in nonracial contexts where conceptual resources are fractured and contested.

Insofar as a multiracial person lives in a hermeneutically unjust conceptual environment, she will be subject to certain characteristic struggles. Racial

[57] Of course, as in Sarah's case, individual interlocutors might treat Diallo unjustly, e.g., by being dismissive, condescending, or inconsiderate.

self-knowledge, in the face of an emaciated conceptual vocabulary, will be more difficult. Communicating one's race to others, including establishing race-based solidarity, may also be challenging. Finally, insofar as (some) multiracial people lack the resources necessary to articulate and communicate their identities, it seems likely that they will also be hindered in their ability to articulate, communicate, and ultimately combat the forms of racial discrimination and racialized harm they experience.

To conclude, I want to offer two remarks. First, although my discussion has focused on racial denials, including the harm that they can inflict on multiracial individuals, I do not mean to suggest that many or most multiracial individuals have mental lives irrepressibly burdened by inchoate and misunderstood racial identities. The pathologization of multiracial identity has a long history, tracing back to the tragic mulatto figure in the 1800s and the Marginal Man hypothesis of the 1900s. Stereotypically, multiracial individuals are portrayed as torn between two worlds, with their mental lives dominated by a tragic sense of fragmentation. For many multiracial writers, it is crucial to replace such stereotypes with a more nuanced understanding of multiracial experience.

The account I have offered may, however, seem to contribute to such stereotypes. In particular, my account implies that multiracial individuals *do* face a challenging hermeneutical environment. However, the claim that multiracial individuals may have more difficulty in racial self-understanding/communication is distinct from the stereotypical claim that multiracial individuals have mental lives marked by a sense of fragmentation. The question of the *significance* of gappy racial hermeneutical environments for one's mental life is, after all, very much dependent upon the person, her circumstances, and what she cares about. For instance, whether the difficulty of articulating one's racial identity dominates one's mental life, and whether one encounters it as *tragic* (as opposed to, for instance, exciting or interesting), will depend on many factors. Some multiracial individuals do grapple, painfully, with racial self-understanding and self-expression. Others do not. For these latter, perhaps, their sense of belonging within the world does not depend so much on racialized self-understanding or the racial acceptance of others.

Indeed, it is worth remembering that even for those individuals who do grapple painfully with questions of racial identity, these questions take their place among many others, and their importance may ebb and flow. The biracial writer Rebecca Walker is one such example. As she willingly attests, she has spent a significant portion of her life with an "unhealthy sense of [racial] fragmentation."[58] As the author of an autobiography on multiracial identity,

[58] Walker, "Introduction," 17.

she has arguably immersed herself in racial questions more deeply than most. Nevertheless, writing six years after the publication of that book, Walker muses:

> I rarely think about being mixed these days, other than to notice the effect it has on others and to consider the assumed implications of it in a racially charged situation. When I do contemplate my mixedness, it is like visiting an old friend, familiar, but no longer involved in the day-to-day goings-on of my life.[59]

That is, we must not lose sight of the banal point that questions of racial self-understanding, no matter how complex or challenging, ultimately constitute just one aspect of a person's life and that the relative importance of racial questions may shift over one's life course.

As a second closing comment, it is worth returning to the larger endeavor that opened this essay—that is, moving multiracial experience from margin to center. How might philosophy of race be enhanced by treating the experiences and understandings of those who are multiracial as central, instead of marginal or exceptional? At the very least, an enhanced focus on multiraciality would raise questions about the relationship between appearance and racial identity (e.g., for multiracial individuals, appearance is not determinative of racial identity), about the role of choice in racial identity (e.g., many multiracial individuals describe a process of exploring different racial identities, raising the possibility that one's racial identity may be partly voluntary), and about the political and pragmatic significance of *declaring* one's racial identity (for monoracial individuals, it is typically not necessary to use speech to claim racial identity since it is assumed on the basis of appearance; in contrast, for multiracial individuals, declarations serve complex political and pragmatic functions). Probing these aspects of multiracial experience has the theoretical potential to deepen our understanding of race. Perhaps more importantly, doing so expresses an ethical commitment to the value of multiracial lives as equally significant within the practice of philosophy.[60]

Occidental College
eroedder@oxy.edu

[59] Walker, "Introduction," 18.

[60] I am grateful to many individuals who have read and commented on this paper, including Cherrish Hardy, Sally Haslanger, Clair Morrissey, Ryan Preston-Roedder, and Dylan Sabo, as well as two anonymous reviewers. Audiences at three workshops offered helpful comments: the International Conference on Identities and Epistemic Justice, the Working Group on Gender and Philosophy, and the Southern California Feminist Philosophy Salon.

REFERENCES

Alba, Richard. *The Great Demographic Illusion: Majority, Minority, and the Expanding American Mainstream*. Princeton, NJ: Princeton University Press, 2020.

Alcoff, Linda Martín. *Visible Identities: Race, Gender, and the Self*. New York: Oxford University Press, 2005.

Anderson, Luvell. "Epistemic Injustice and the Philosophy of Race." In *The Routledge Handbook of Epistemic Injustice*, edited by Ian James Kidd, José Medina, and Gaile Pohlhaus Jr., 139–48. New York: Routledge, 2017.

Anzaldúa, Gloria E. *Borderlands/La Frontera: The New Mestiza*. San Francisco: Aunt Lute Books, 1987.

Appiah, Kwame Anthony. *The Ethics of Identity*. Princeton, NJ: Princeton University Press, 2005.

Appiah, Kwame Anthony, and Amy Gutmann. "Race, Culture, Identity: Misunderstood Connections." In *Color Conscious: The Political Morality of Race*, 30–105. Princeton, NJ: Princeton University Press, 1998.

Barnes, Elizabeth. *The Minority Body: A Theory of Disability*. Oxford: Oxford University Press, 2016.

Botts, Tina Fernandes, ed. *Philosophy and the Mixed Race Experience.* Lexington Books, 2016.

Collins, Patricia Hill. "Learning from the Outsider Within: The Sociological Significance of Black Feminist Thought." *Social Problems* 33, no. 6 (October–December 1986): S14–S32.

Daniel, G. Reginald. *More Than Black: Multiracial Identity and New Racial Order*. Philadelphia: Temple University Press, 2002.

Daniel, G. Reginald, Laura Kina, Wei Ming Dariotis, and Camilla Fojas. "Emerging Paradigms in Critical Mixed Race Studies." *Journal of Critical Mixed Race Studies* 1, no. 1 (2014): 6–65.

Davenport, Lauren. "The Fluidity of Racial Classifications." *Annual Review of Political Science* 23 (May 2020): 221–40.

———. *Politics Beyond Black and White: Biracial Identity and Attitudes in America*. Cambridge: Cambridge University Press, 2018.

Davis, F. James. *Who Is Black? One Nation's Definition*. 10th Anniversary ed. University Park, PA: Pennsylvania State University Press, 2001.

Degler, Carl N. *Neither Black nor White: Slavery and Race Relations in Brazil and the United States.* New York: MacMillan, 1971.

Dotson, Kristie. "A Cautionary Tale: On Limiting Epistemic Oppression." *Frontiers: A Journal of Women Studies* 33, no. 1 (January 2012): 24–47.

Elam, Michele. *The Souls of Mixed Folk: Race, Politics, and Aesthetics in the New*

Millennium. Stanford, CA: Stanford University Press, 2011.

Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press, 2007.

Grillo, Trina. "Anti-Essentialism and Intersectionality: Tools to Dismantle the Master's House." *Berkeley Women's Law Journal* 10 (1995): 16–30.

Guevarra, Rudy P., Jr. *Becoming Mexipino: Multiethnic Identities and Communities in San Diego*. New Brunswick, NJ: Rutgers University Press, 2012.

Hall, C. C. I. "Please Choose One: Ethnic Identity Choices for Biracials." In *Racially Mixed People in America*, 250–64. Newbury Park, CA: Sage, 1992.

Hardimon, Michael O. "The Ordinary Concept of Race." *Journal of Philosophy* 100, no. 9 (September 2003): 437–55.

Harris, Angela P. "Race and Essentialism in Feminist Legal Theory." *Stanford Law Review* 42, no. 3 (February 1990): 581–616.

Harris, Melvin. *Patterns of Race in the Americas*. New York: W. W. Norton, 1964.

Haslanger, Sally. "Gender and Race: (What) Are They? (What) Do We Want Them To Be?" *Noûs* 34, no. 1 (March 2000): 31–55.

———. *Resisting Reality: Social Construction and Social Critique*. New York: Oxford University Press, 2012.

———. "You Mixed? Racial Identity without Racial Biology." In *Adoption Matters: Philosophical and Feminist Essays*, edited by Sally Haslanger and Charlotte Witt, 265–90. Ithaca, NY: Cornell University Press, 2006.

Hitlin, Steven, Glen H. Elder Jr., and J. Scott Brown. "Racial Self-Categorization in Adolescence: Multiracial Development and Social Pathways." *Child Development* 77, no. 5 (September–October 2006): 1298–308.

Jeffers, Chike. "Cultural Constructionism." In *What Is Race? Four Philosophical Views*, by Joshua Glasgow, Sally Haslanger, Chike Jeffers, and Quayshawn Spencer, 38–72. New York: Oxford University Press, 2019.

Johnston, Marc P., and Kevin L. Nadal. "Multiracial Microaggressions: Exposing Monoracism in Everyday Life and Clinical Practice." In *Microaggressions and Marginality: Manifestation, Dynamics, and Impact*, edited by Derald Wing Sue, 123–44. New York: John Wiley and Sons, 2010.

Jordan, Winthrop D., and Paul Spickard. "Historical Origins of the One-Drop Racial Rule in the United States." *Journal of Critical Mixed Race Studies* 1, no. 1 (2014): 98–132.

Khanna, Nikki. *Biracial in America: Forming and Performing Racial Identity*. Lanham, MD: Lexington Books, 2011.

———. "Ethnicity and Race as 'Symbolic': The Use of Ethnic and Racial Symbols in Asserting a Biracial Identity." *Ethnic and Racial Studies* 34, no. 6 (2011): 1049–67.

Kitcher, Philip. "Race, Ethnicity, Biology, Culture." In *Racism: Key Concepts* *in Critical Theory*, edited by Leonard Harris, 87–117. New York: Humanity Books, 1999.

Lindemann, Hilde. *Holding and Letting Go: The Social Practice of Personal Identities*. New York: Oxford University Press, 2014.

MacIntyre, Alasdair C. *After Virtue: A Study in Moral Theory*. Notre Dame, IN: University of Notre Dame Press, 1981.

McKibbin, Molly Littlewood. *Shades of Gray: Writing the New American Multiracialism*. Lincoln, NE: University of Nebraska Press, 2018.

McPherson, Lionel K. "Deflating 'Race.'" *Journal of the American Philosophical Association* 1, no. 4 (Winter 2015): 674–93.

Mills, Charles W. "'But What Are You Really?' The Metaphysics of Race." In *Blackness Visible: Essays on Philosophy and Race*, edited by Charles W. Mills, 41–66. Ithaca, NY: Cornell University Press, 1998.

Pohlhaus, Gaile, Jr. "Relational Knowing and Epistemic Injustice: Toward a Theory of Willful Hermeneutical Ignorance." *Hypatia* 27, no. 4 (Fall 2012): 715–35.

Ratliff, Sarah, and Bryony Sutherland. *Being Biracial: Where Our Secret Worlds Collide*. Collinsville, MS: Heritage Press, 2015.

Renn, Kristen A. *Mixed Race Students in College: The Ecology of Race, Identity, and Community on Campus*. Albany, NY: State University of New York Press, 2004.

———. "Research on Biracial and Multiracial Identity Development: Overview and Synthesis." *New Directions for Student Services* 2008, no. 123 (Autumn 2008): 13–21.

Rockquemore, Kerry Ann, and David L. Brunsma. *Beyond Black: Biracial Identity in America*. 2nd ed. Lanham, MD: Rowman and Littlefield Publishers, 2007.

Root, Maria P. P. "The Multiracial Experience: Racial Borders as a Significant Frontier in Race Relations." In *The Multiracial Experience: Racial Borders as the New Frontier*, xiii–xxviii. Thousand Oaks, CA: Sage, 1996.

Schechtman, Marya. *The Constitution of Selves*. Ithaca, NY: Cornell University Press, 1996.

Schwartz, Lacey, dir. *Little White Lie*. Truth Aid and ITVS, 2014.

Song, Miri. "Who Counts as Multiracial?" *Ethnic and Racial Studies* 44, no. 8 (2021): 1296–323.

Spencer, Quayshawn. "A Radical Solution to the Race Problem." *Philosophy of Science* 81, no. 5 (December 2014): 1025–38.

Strmic-Pawl, Hephzibah V. *Multiracialism and Its Discontents: A Comparative Analysis of Asian-White and Black-White Multiracials*. Lanham, MD: Lexington Books, 2016.

Sundstrom, Ronald R. "Being and Being Mixed Race." *Social Theory and Practice* 27, no. 2 (April 2001): 285–307.

———. *The Browning of America and the Evasion of Social Justice*. Albany, NY: State University of New York Press, 2008.

Taylor, Charles. *Sources of the Self: Making of the Modern Identity*. Cambridge, MA: Harvard University Press, 1989.

Townsend, Sarah S. M., Hazel R. Markus, and Hilary B. Bergsieker. "My Choice, Your Categories: The Denial of Multiracial Identities." *Journal of Social Issues* 65, no. 1 (March 2009): 185–204.

Walker, Rebecca. *Black, White, and Jewish: Autobiography of a Shifting Self*. New York: Riverhead Books, 2002.

———. "Introduction." In *Mixed: An Anthology of Short Fiction on the Multiracial Experience*, edited by Chandra Prasad. New York: W. W. Norton, 2006.

Williams, Nathaniel Andrew, and Caroline Ware. "A Tale of Two 'Halfs': Being Black, While Being Biracial." *International Journal of Qualitative Studies in Education* 32, no. 1 (2019): 85–106.

Zack, Naomi. *Race and Mixed Race*. Philadelphia: Temple University Press, 1993.

———. "How Mixed Race Is Not Constructed: US Identities and Perspectives." In *The Oxford Handbook of Philosophy and Race*, edited by Naomi Zack. New York: Oxford University Press, 2017.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 2 · MARCH 2024

https://doi.org/10.26556/jesp.v27i1.2161
© 2024 Author

IMMORAL ARTISTRY?

REFLECTIONS ON OMAR LITTLE, TONY SOPRANO, AND VALUE INTERACTION DEBATES

Sam Shpall

THE RELATIONSHIP between moral and aesthetic value has preoccupied philosophers at least since Plato and animated many influential treatments of art criticism since David Hume's "Of the Standard of Taste." Contemporary philosophers have been especially interested in the status of *autonomism*—the view that an artwork's moral virtue or vice is irrelevant to its aesthetic value—and in a variety of nonautonomist positions.[1] Here I will be exploring the controversial position known as *immoralism*. According to immoralists, moral flaws can make a positive contribution to a work's aesthetic value.[2]

I hope to provide a distinctive perspective on debates about immoralism and value interaction more generally, grounding this perspective in philosophically engaged art criticism. Unlike many other writers on these topics, I am not primarily interested in whether immoralism is true. Immoralism is supposed to provide an answer to a philosophically significant question about the interaction of values. Yet the standard formulations of this question are frustratingly ambiguous. Can a moral flaw make an artwork better aesthetically? In order to evaluate this question, we must ask another, that is, "better than what?"

Consider the oft-rehashed case of *Triumph of the Will*.[3] According to Daniel Jacobson, following Susan Sontag, the film's moral defects are "inseparable"

[1] I am not concerned about the distinction between artistic and aesthetic value in this paper and will employ the notions interchangeably.

[2] Several formulations of this thesis appear in the literature. Panos Paris understands immoralism as the view that an artwork can be aesthetically better in virtue of its immorality ("The 'Moralism' in Immoralism"). Moonyoung Song argues that we should understand it as the view that a moral defect can itself be an aesthetic merit ("The Nature of the Interaction between Moral and Artistic Value").

[3] I agree with Rafe McGregor that *The Birth of a Nation* is a more compelling example ("A Critique of the Value Interaction Debate," 462). But my point in the text is that this does not matter. All examples like this are of limited interest.

from its aesthetic value: its political and aesthetic ideals are unified.[4] Berys Gaut claims, similarly, that the film is held together by its offensive celebration of Nazism.[5] We seem to be off and running. If we were to extricate the Nazism, the film would be morally better. But it would also be aesthetically worse, since a sanitized *Triumph* would be incoherent at best. So, its moral defects make the film better as art. Immoralism is true.[6]

I am not impressed. *Triumph* could not exist without its Nazism![7] That is why the claim that the work would have been aesthetically better if it had vilified rather than glorified Nazism is puzzling. Jacobson says this claim is "either meaningless or false."[8] I am happy to admit that *Triumph* has more aesthetic value than no film at all. But that is not an energizing comparison, and it does not do much to support immoralism.[9]

Central examples from the literature on comic immoralism are awkward for the same reason. Here is the main shtick of Sacha Baron Cohen's *Da Ali G Show*: the comedian manipulates people into embarrassing revelations on camera via lies about his identity. Suppose the manipulation is immoral and the revelations are funny. Argument: without the manipulation, the show would be morally better but comically worse. So, (comic) immoralism is true.[10]

Again, this is inconclusive. *Da Ali G Show* would be totally unrecognizable without Baron Cohen's manipulative methods; the methods are an essential condition of the work's existence. What could we mean in claiming that his immorality makes the work comically better? Better than what? We certainly cannot say: "A completely different artwork he might have made instead"![11]

[4] Jacobson, "In Praise of Immoral Art," 192; Sontag, "Fascinating Fascism." See also Jacobson, "Ethical Criticism and the Vice of Moderation"; and Kieran, "Forbidden Knowledge."

[5] Gaut, *Art, Emotion, and Ethics*, 190.

[6] John, "Artistic Value and Opportunistic Moralism"; and Stear, "Immoralism Is Obviously True."

[7] Stecker, "Immoralism and the Anti-Theoretical View"; Song, "The Nature of the Interaction between Moral and Artistic Value."

[8] Jacobson, "In Praise of Immoral Art," 193.

[9] See Kieran, "Art, Imagination, and the Cultivation of Morals," for further discussion of *Triumph*.

[10] Nannicelli, "Moderate Comic Immoralism and the Genetic Approach to the Ethical Criticism of Art." I think comic immoralism entails immoralism about aesthetic value more generally, because comically valuable properties are often aesthetically valuable. Nothing here will depend on that contention. Nils-Hennes Stear gives a noncomic example with a similar structure: photographer Jeff Mermelstein's *#nyc* series, which captures people's intimate text messages without their consent ("Immoralism Is Obviously True").

[11] Some fans of Baron Cohen will deny that his deceptive practices are immoral. Maybe his deception is *prima facie* wrong, but ultimately justified because of the socially valuable

It would probably be more useful to appeal to cases in which transgressions are less global, where the immorality is not so essential to the work's identity. The "Better than what?" question might have, in such examples, the following answer: better than the sufficiently similar version of that artwork would have been were the immoral feature removed. For the sake of argument, I assume there are such examples. That is because my aim is to pursue a slightly different question, one that I think is much more central to the practice of art interpretation and criticism than the question of whether immoralism is true. Supposing that moral flaws can make artworks better aesthetically, how can they do this? More specifically: Are there general strategies that artists can, do, and should (sometimes) pursue to exploit immorality for aesthetic ends?

I have explained why I am not interested in the answer that is implicit in the *Triumph* example and many others like it, which is that moral flaws can make artworks "better" aesthetically when the artwork could not exist without them. Ditto for the related answer that is implicit in some treatments of comic immoralism, which is that comedy of some valuable forms requires cruelty, or deception, or other moral transgressions. These claims might be true, but they do not amount to illuminating characterizations of artistic strategies. After all, everyone admits that some racist artworks are bad as artworks and no better for their racism. Everyone, comic immoralists included, admits that cruelty often backfires comically. Even if we agree that racism can produce artistic value or that cruelty can produce comic value, it is reasonable to be curious about when and how they do so. Gaut's convincing discussion of artistic strategies shows how artists can deliver moral understanding in aesthetically valuable ways.[12] My guiding question is whether his immoralist opponents can provide similarly compelling conceptualizations of immoral artistry.

The discussion is structured around the best attempt to outline an aesthetically productive immoralist strategy. We find this attempt in the work of A. W. Eaton.[13] Eaton examines what she takes to be a widely employed artistic strategy involving a distinctive character type (the "rough hero"). She gives a fascinating and provocative argument for immoralism on the basis of the achievements of rough hero works. Though this argument has been discussed by a number of philosophers of art, I will be drawing out several themes that deserve more attention, stressing the ways that value interaction debates can be usefully connected to broader moral psychological inquiry.

revelations it prompts. See Nannicelli, "Moderate Comic Immoralism and the Genetic Approach to the Ethical Criticism of Art," 174.

[12] Gaut, *Art, Emotion, and Ethics*, 186–94.

[13] Eaton, "Robust Immoralism" and "Reply to Carroll."

I will also be exploring the philosophical significance of detailed art interpretation and its relationship to the methodology of aesthetics. In discussions of comic value, immoralists have at times recognized the difficulties with popular example-based arguments they would like to endorse.[14] A similar dynamic complicates the evaluation of favored immoralist examples from the history of literature and film, as I have already suggested. These examples are often "too messy to be effective."[15] Even philosophers committed to immoralism tend to recognize that many such arguments are inconclusive.[16]

My interpretations of *The Wire* and especially *The Sopranos* aim to convince readers that various moves in value interaction debates have presupposed misguided readings of the artworks invoked as evidence. The main goal is not to vindicate a final judgment on immoralism itself but to facilitate exploration of the question about artistic strategies via philosophical interpretation that sensitively engages with the relevant artworks—as well as the insightful art criticism about them that already exists and is seldom written by philosophers. In addition to expressing a point of view on the way we use examples in aesthetic theorizing, I aim to contribute to critical appreciation of these works, *The Sopranos* in particular. Fans of the show will have to judge whether my interpretation is at all original and whether the philosophical backdrop contributes to a convincing critical appraisal.

1. Overview of Eaton's Argument and My Critique

Here is my understanding of Eaton's argument:

1. The rough hero is irredeemably vicious.[17]
2. It is morally bad to sympathize with an irredeemably vicious character.[18]

[14] Ted Nannicelli writes: "In the context of comedy, at least, it is rarely the case that a work actually endorses the immoral behaviors that it represents" ("Moderate Comic Immoralism and the Genetic Approach to the Ethical Criticism of Art," 171).

[15] Stecker, "Immoralism and the Anti-Theoretical View," 157 (commenting on John, "Artistic Value and Opportunistic Moralism").

[16] Compare Li, "Immorality and Transgressive Art."

[17] This is a simplification of Eaton's argument in "Robust Immoralism" (284), where she claims that the hero is (a) grievously flawed, (b) flawed at the level of deep character, (c) remorseless, and (d) lacks virtues sufficient to outweigh his flaws. I use "irredeemable viciousness" as a term of art that stands in for this account. I discuss various difficulties below.

[18] This too is a simplification. Eaton also worries about our *endorsement of*, *admiration for*, and *siding with* the rough hero. In the present formalization I state the relevant claims as concisely as possible. My later formulations will remind readers of this crucial ambiguity.

3. So, it is morally bad to sympathize with the rough hero.
4. Rough hero works encourage us to sympathize with the rough hero.
5. So, rough hero works encourage us to do something morally bad.
6. Getting us to sympathize with the rough hero is an aesthetic achievement.
7. So, rough hero works encourage us to do something morally bad, and if they get us to do this thing, it is an aesthetic achievement.[19]

Alternatively: sympathy for the rough hero is morally bad, so if an artwork encourages us to have this mental state, it is in that respect morally bad. But successfully getting us to feel sympathy is aesthetically good, because it represents an interesting, indeed "delicious," overcoming of imaginative resistance.[20] Immoralism is true.

I will note one subtlety before we move to the fun stuff. For Eaton, a character's being irredeemably vicious *implies* that it is wrong to sympathize with them—that is, wrong to like them, admire them, and root for them (premise 2). Indeed, Eaton's definition of the rough hero genre invokes two conditions that embody this connection and explain the supposed immorality of the relevant works. First, the rough hero is irredeemably vicious. Second, the work presents them sympathetically.[21] In other words, the work's sympathetic presentation of an irredeemably vicious character is what makes it morally flawed.

By contrast, I think it is important to sharply distinguish claims about the viciousness of characters from claims about how we are morally required to respond to them. And it is important to distinguish both from claims about whether an artwork encourages or prescribes specific reactions to its characters. The structure of my formulation of the above argument reflects this division, as does the structure of my critique in the rest of the paper. I pursue a response to Eaton that is best understood disjunctively. My suspicion is that most or all of Eaton's examples have at least one of the following properties:

A. The relevant character is not irredeemably vicious (*contra* premise 1).

[19] Encouraging us to sympathize is not the same thing as getting us to sympathize. So, according to Eaton, the moral defect is in one sense prior to the aesthetic merit. I do not share Song's judgment that this means the work can only be aesthetically valuable despite its moral defect rather than in virtue of it ("The Nature of the Interaction between Moral and Artistic Value," 292).

[20] Eaton, "Robust Immoralism," 287. See Tamar Gendler's "The Puzzle of Imaginative Resistance" for a classic discussion of imaginative resistance.

[21] For the latter condition, see Eaton, "Robust Immoralism": "Although the rough hero is supposed to be morally hateful, he is also supposed to be a *hero*; that is, a sympathetic, likeable, and admirable protagonist" (285).

B. It is not morally bad to sympathize with the relevant character in the relevant way (*contra* premise 2).

C. The artwork does not encourage the relevant form of sympathy (*contra* premise 4).

Finally, some of the best responses to my arguments will commit proponents of immoralism to the view that getting us to sympathize with the rough hero is no special achievement at all (*contra* premise 6).

2. OMAR LITTLE AND IRREDEEMABLE VICE

My first main claim is that some of Eaton's "rough heroes" are not rough heroes. They are not rough heroes because they are not irredeemably vicious. It would take several books to adequately explore all the fictional works Eaton mentions, so I will concentrate on one specific case, the character of Omar Little in the phenomenal television series *The Wire*. I focus on this example because the erroneous categorization of Omar as a rough hero is particularly suggestive.

Eaton regards Omar as a "glorified criminal."[22] She briefly defends this judgment by saying that "while Omar adheres to a strict code, his criminal activity is always aimed solely at promoting his own good rather than taking law enforcement into his own hands."[23] This is not true. Omar pursues the good of various others in addition to his own and a brand of justice to the detriment of his own good. More importantly, the characterization of Omar as a rough hero on these grounds is unconvincing.

It gives me a peculiar delight to stick up for Mr. Little, one of the most beloved characters in modern American television—a favorite of critics, most of the show's viewers, and even President Obama. Omar inspires these affections because he resists easy categorization and evaluation. He glitters with moral complexity and invites question as much as judgment. The complexity that attracts many of us to Omar is incompatible with irredeemable immorality as Eaton understands it.

Omar is a freelance bandit who makes his living stealing from violent drug kingpins. He does so with a "splendidly and improbably diverse troop of soldiers," including women, fellow members of sexual minorities, elderly former gangsters, and a blind man.[24] Omar's courage and cunning make him an object

[22] Other examples include Bonnie and Clyde from *Bonnie and Clyde*, Michael Corleone from *The Godfather*, Gus Fring from *Breaking Bad*, William Munny from *Unforgiven*, Vincent and Jules from *Pulp Fiction*, and her paradigm case, Tony Soprano from *The Sopranos*.

[23] Eaton, "Reply to Carroll," 380n12.

[24] Cormier, "Bringing Omar Back to Life," 210.

of *sui generis* admiration. His legendary exploits, spectral materializations, and ridiculous chutzpah are singularly capable of inspiring fear, even in the most hardened criminals.[25] His nemesis Marlo Stanfield revealingly compares him to Spider-Man. Further, Omar thinks of his victims as evil or at least as undeserving of their money. The latter thought is not misguided. In sum, he is uniquely and attractively threatening to the corrosive institutions of the criminal underworld.

Despite these capacities, and unlike many of *The Wire*'s villainous characters, Omar carefully avoids gratuitous violence and illegality. He does not deploy his considerable criminal talents beyond the special sphere of extorting traffickers and dealers. "A man gotta have a code," he says. "Don't get it twisted, I do some dirt too, but I never put my gun on no one who wasn't in the game." He refuses to snitch unless there is just cause. When he does collaborate with law enforcement, he does so not primarily because it serves his financial interests but because he is morally interested in punishing the most indiscriminately violent members of Baltimore's drug trade—when, and only when, they have violated what he sees as the rules of the game. This scrupulousness is laudable notwithstanding the idiosyncrasy of his moral code, which also involves forswearing curse words, regularly taking his grandmother to church in a taxi, and observing all gang truces religiously.

When he wants to be, of course, Omar is a brutal executioner. But even brutal executioners can have good qualities. He is bold and streetwise, cool and meditative, laconic and witty, fearless and determined. From his iconic showdown with another likable criminal, Brother Mouzone: "This range? This caliber? Even if I miss I can't miss."

Omar also exhibits compassion and loving-kindness. He is sensitive and affectionate with many acquaintances, including some police officers. He is even more sensitive and loving with friends and romantic partners. Our justified sympathy for Omar intensifies when his lover Brandon is tortured, mutilated, and killed by Avon Barksdale's enforcers. Though the risk of death is "all in the game," this kind of treatment is not. The experience makes Omar's desire for retribution understandable. A similar desire brings him out of retirement years later and precipitates his downfall: he only returns to Baltimore from Puerto Rico to avenge the torture and murder of his old friend Butchie.

Finally, Omar is an openly gay Black man living on the margins of an intensely heteronormative culture. He represents queer masculinity unapologetically in a particularly hostile context. This courageous pride moderates our condemnation. Thematically speaking, the bitter homophobia occasioned by Omar's sexuality dramatizes deep questions about how heteronormative

[25] Shuster, *New Television*, 108.

patriarchy contributes to "toxic masculinity," male deviancy, and urban decay.[26] To its credit, *The Wire* positions homophobia as a social ill afflicting the police as much as boys on the corner, intensifying our appreciation of Omar's radical challenge to damaging stereotypes of masculinity.

None of this means Omar Little is a paragon of virtue. *The Wire* explicitly disavows any triumphalist interpretation of his character. In a celebrated scene, Detective Bunk Moreland eloquently condemns Omar's callousness and his complicity in poisonous structures of violence. Bunk: "As rough as that neighborhood could be, we had us a community . . . nobody, no victim who didn't matter . . . and now all we got is bodies . . . and predatory motherfuckers like you. . . . Out where that girl fell, I saw kids acting like Omar, calling you by name, glorifying your ass . . . makes me sick motherfucker how far we done fell." That this speech expresses part of *The Wire*'s perspective on Omar is confirmed by the stickup man's reaction. As Bunk walks away in disgust, a tear rolls down Omar's cheek, and he spits in an ambiguous gesture of rejection. Bunk turns around to look at him. Omar's spit hangs on his chin, as if confirming that Bunk's moral force has overcome his attempt to dismiss it. This "rough hero" is not rough enough to raise his eyes and meet the detective's.

Omar exhibits grave moral defects alongside admirable moral virtues. He exhibits nonmoral virtues—streetwise intelligence, physical dexterity, coolness, wit, style—that are hard to weigh against moral ones in any sort of definitive evaluation. He appears to exhibit deep remorse. Whether Omar's virtues ultimately outweigh his flaws is a question that I find odd, and that I will address at a general level momentarily. In any event, no viewer of *The Wire* could conclude that he is on a moral par with the villainous Stanfield or the more unequivocally malevolent of Eaton's glorified criminals. For these reasons, he is a poor candidate for the sort of "morally hateful" protagonist needed to underwrite the argument for immoralism.

Here is a transitional observation. I have been using the notion of irredeemable vice to stand in for the more expansive set of properties Eaton employs to conceptualize the rough hero: being grievously flawed, remorseless, and "more bad than good."[27] However, these are coarse-grained evaluations, whose

[26] I use "deviancy" here in a nonmoralized way to refer to illegal and often imprudent behavior. One of the most compelling features of *The Wire* is its extended presentation of the idea that moral corruption is only one causal ingredient in patterns of deviant behavior. Compare Shelby, "Justice, Deviance, and the Dark Ghetto."

[27] I ignore Eaton's additional "deep character" condition. Genuinely grievous flaws must be deep character flaws. If a flaw is superficial or not a matter of the agent's character, then it is not grievous enough to contribute significantly to irredeemable viciousness. See Eaton, "Robust Immoralism," 284.

relations to distinctive prohibitions on sympathetic reactions are highly contestable. First, deep character flaws regularly cohabit with virtues, and it would take some argument to motivate the idea that such flaws make sympathetic reactions to virtues or the persons possessing them morally suspect. Second, the moral value of remorse is a topic of legitimate philosophical disagreement, and remorsefulness is only one kind of appropriate reparation for immoral behavior—which is sometimes unnecessary for redemption and often insufficient for it. Third, though we sometimes judge that people or characters are *bad overall*, most of us do this relatively rarely and only in egregious circumstances of immoral behavior. It is possible that we are in general too quick to do this.

This is not to say that there are no evil people or evil characters. It is to express skepticism about the ease of identifying irredeemable vice. The observation may seem unimportant given my admission that some rough characters are likely irredeemably vicious on any reasonable understanding. But this misunderstands the shape of the critique. The point of the observation is to remind us that most interesting fictional characters are, like Omar Little and many real people, complex mixtures of good and bad, inviting appropriately ambivalent reactions, including some appropriately sympathetic ones.[28] I will now discuss the subtlety of these appropriateness conditions in more detail.

3. TONY SOPRANO AND SYMPATHIZING WITH EVIL CHARACTERS

I will now argue that Eaton oversimplifies the nature of character evaluation. Taking for granted that there are some irredeemably vicious characters in fiction, I will cast doubt on the view that it is morally bad to sympathize with them. This skepticism expresses a general perspective on moral evaluation that has implications beyond the value interaction debate and the sphere of art appreciation, though its implications in the context of art are distinctive.

The reader will have gathered that I may not have much of a handle on the notion of irredeemable vice. Nonetheless, I can recognize some plausible candidates. Consider the real-life killer Robert Alton Harris, who brutally murdered two teenage boys and was executed in 1992 in San Quentin State Prison,

[28] It seems to me indicative, for instance, that Fyodor Dostoevsky's "underground man" (a rough hero, for Eaton) is a far more interesting figure than most more unambiguously villainous characters. Usually, artistic prescriptions and actual audience reactions track these differences. We are encouraged to sympathize with the underground man in unique ways, we do tend to sympathize in these ways, and this is morally appropriate. See Richard Pevear's foreword to *Notes from Underground* for discussion of this complicated protagonist.

California. In the leadup to his execution, fellow inmates on death row pledged money for a candy and soda party so they could celebrate his demise.[29]

Gary Watson analyzes the life of this reviled killer at length in a brilliant and famous essay. An arresting feature of the discussion is that it encourages us to sympathize with Harris, largely on account of his abominable upbringing. Successfully encouraging this sympathetic reaction is an essential feature of Watson's argument about the nature of reactive attitudes—as is the claim that this sympathy is *appropriate*. It is appropriate sympathy for Harris that explains our confused and unstable responsibility judgments, and this is the fact about reactive attitudes that Watson aims to explore.[30]

It is hard to believe that Watson is doing something immoral in provoking sympathy for Harris. It is better to say what Watson says: that some forms of sympathy for evil people are permissible and even appropriate or laudable. Other forms of sympathy for such people are morally unacceptable. Sympathy with Harris on account of his terrible childhood is compatible with antipathy toward his behavior and his character as an adult.

Why not tell a similar story about sympathizing with irredeemably vicious fictional characters? For the moment I set aside the fact that a character's fictionality itself has serious implications for what forms of response are possible and appropriate, though I will return to this below. Consider Eaton's chief example of a rough hero, the mob boss Tony Soprano from *The Sopranos*. When we say that we like or admire Tony, one reading of this thought is that we "compartmentalize" our sympathy: we like or admire some things about Tony while being repulsed by other aspects of his character. Sensitive viewers are attuned to his faults just as they are attuned to his charms. One of the main joys of engaging with the series is becoming invested in this jumble of reactions.

This perspective is common.[31] I will respond to some objections to it later in this section. But first, I want to begin providing a substantive interpretation

[29] Miles Corwin, "Icy Killer's Life Steeped in Violence":

> "The guy's a misery, a total scumbag; we're going to party when he goes," said Richard (Chic) Mroczko, who lived in the cell next to Harris on San Quentin Prison's Death Row for more than a year. "He doesn't care about life, he doesn't care about others, he doesn't care about himself."

[30] Watson, "Responsibility and the Limits of Evil":

> What appears to happen is that we are unable to command an overall view of his life that permits the reactive attitudes to be sustained without ambivalence … in light of the 'whole' story, conflicting responses are evoked. The sympathy toward the boy he was is at odds with outrage toward the man he is … each of these responses is appropriate. (244)

[31] See Carroll, "Rough Heroes": "One can admire Tony's attempts to be a good father … without morally endorsing Tony's garroting squealers" (373). Paris asks: "Would not a work

that can help motivate and develop the claim about compartmentalized sympathy for evil characters. As I hope is clear, "compartmentalized sympathy" is shorthand for the compartmentalization of a large set of reactive attitudes and emotional responses.

Though Tony is Eaton's motivating example, and though many philosophers have discussed Eaton's argument and Tony's role in it, the literature on immoralism does not often engage with the large body of critical work on *The Sopranos*. That is unfortunate, because many writers have addressed moral objections to the show in sophisticated ways. It is worth remembering that two decades ago, this series was the topic of extraordinary public attention, prompting denunciations from conservative writers, Republican politicians, feminist media critics, and various Italian American organizations—as well as awed praise from prominent film theorists, crime reporters, psychotherapists, and even the real-life Donnie Brasco.[32] Sensitively contextualizing longstanding debates about *The Sopranos* must be part of any serious moral reckoning with it. Similarly, thorough interpretation is the inevitable groundwork for theoretical arguments that turn on claims about its ambitions and achievements.

In this section and the next, I will explain what philosophers invoking the example of Tony Soprano have tended to overlook. In this section, I will focus on the surprising difficulties we encounter in providing a succinct evaluation of Tony's moral character and the reactive attitudes it ought to occasion. In the next section, I will say more about why it is unfair to convict *The Sopranos* of encouraging us to sympathize with its central character in immoral ways. My broader contention is that elements of the perspective defended here likely generalize to other examples of supposed immoral artistry, though I cannot make good on that claim in this paper.

Let us begin with a comparison. On the face of it, Tony Soprano is much worse than Omar Little. We have discussed Omar's code and his sincere commitment to honoring it. Tony similarly thinks of himself as a scrupulous soldier, a champion of "family" and "honor," but he prioritizes "business" in a way that reveals his rhetoric to be little more than narcissistic grandstanding. After all, he unhesitatingly eliminates all obstacles to his criminal enterprising even if they belong to the family—with efficient, remorseless brutality. He orders the killings of Big Pussy and Adriana as soon as he knows the risks they pose. He executes his nephew and protégé, Christopher, himself.

that prescribed dislike for [such characters] through and through, and that took no heed of their positive qualities, be likewise immoral, shallow, or hypocritically moralistic?" ("The 'Moralism' in Immoralism," 21).

[32] For discussion, see Lavery, "'Coming Heavy'"; and O'Brien, "A Northern New Jersey of the Mind."

Tony is also unapologetically sexist, racist, and homophobic. What is more, he is outrageously hypocritical about these prejudices.[33] He justifies his anti-Black racism by invoking Black criminality and parades a zealous fatphobia despite self-identifying as a "fat fucking crook from New Jersey."

As this hypocrisy suggests, Tony's cruelty results from a profoundly stunted moral character. His constant fat shaming most obviously reveals his "rage turned inward," which is how his therapist Dr. Jennifer Melfi evocatively characterizes his depression. He is pathologically incapable of engaging with difficult emotions. His wife Carmela calls him a "wall" because of his expertise in deploying the silent treatment. Of his preferred emotional modes, silence is at least less scary than rage. Even Tony's experience of positive emotions is diseased. He cries over his dead horse Pie-O-My but does not spare a thought for Adriana. As Melfi observes in the wake of the horse's death: "The only other time you've been this emotional in here was for the ducks. You haven't grieved for your mother or other human beings."

So, all reasonable viewers agree that Tony is an angry, violent, vulgar, hateful, duplicitous, callous, self-hating, alexithymic, fatphobic, racist, misogynistic, manipulative, entitled, sociopathic extortionist and murderer.[34] But this is just the beginning of the story! Notwithstanding these abominable characteristics, many viewers feel affection for Tony.[35] I think there are two main sources of this affection and our resulting fascination. First, Tony has admirable qualities. Second, we learn about his terrible moral qualities while also learning many judgment-complicating facts about his moral formation.

Many critics take the core accomplishment of *The Sopranos* to be its convincing juxtaposition of the mobster genre with the kind of suburban domestic drama more commonly associated with soap operas and sitcoms.[36] Central to

[33] Baldanzi, "Bloodlust for the Common Man," 86. Tony is just as hypocritical about the value of work and community, waxing poetic about the church built by his grandparents while scamming the Department of Housing and Urban Development (Polan, "*The Sopranos*," 139), selling out his longtime poultry shop tenants to collect Jamba Juice dollars, etc.

[34] Melfi puts it more simply in response to his stalking her (and demanding her reasons for rejecting his advances!): "Well … you're not a truthful person. You're not respectful of women. You're not really respectful of people.… You take what you want from them by force, or the threat of force." Tony's response: "Fuck you! You fucking cunt!"

[35] Carroll, "Sympathy for the Devil"; Harold, "A Moral Never-Never Land"; and Eaton, "Robust Immoralism."

[36] Most obviously, the technique of "crosscutting" (made famous in *The Godfather*) explicitly associates scenes of extreme violence and domestic warmth (Holden, introduction to *The New York Times on "The Sopranos,"* xiii). According to Geoffrey O'Brien, Tony represents the "domesticated end point for the romance of gangsterism that looks to be America's most durable contribution to world folklore" ("A Northern New Jersey of the Mind," 167). For

this conceit is Tony's authentic investment in the life of his immediate family.[37] He experiences deep, believable parental love for his children, Meadow and A. J. He is committed to becoming better as a father and apologizes sincerely when he lets them down. He cares about the friends and acquaintances of his children (at least the Caucasian ones). He listens to Carmela when it concerns his children's well-being, restraining his otherwise notoriously ungovernable impulses. And though he is a prodigious philanderer, his marital relationship is in other respects surprisingly respectful. Ellen Willis calls Carmela Tony's "emotional equal."[38] Cindy Donatelli and Sharon Alward say that Carmela, his mother Livia, and his sister Janice "have him by the balls, and he knows it."[39] These claims might be overstated, but it is plausible that Tony cares deeply about Carmela, as he cares deeply about his children and perhaps some other family members, and that we identify with his genuine attachment to the value of family.[40] These features of Tony's psychology fascinate because they are so jarringly incongruent with the cruelty he exhibits outside the home.

Tony is also loved by his children, by Carmela, and (at least on the face of it) by many of his friends and associates. This has obvious consequences for our sympathies: when decent people love someone, this licenses at least some justifiable hesitation about, for example, gleefully rooting for their death. Carmela is a complicated, compromised character, but I think it is appropriate to sympathize with her in various ways. Adriana is also compromised, but it is impossible not to sympathize with her, as successive FBI agents setting out to manipulate her discover. Both these characters love Tony; their love is grounded in value judgments that are skewed yet comprehensible. We have no reason to doubt the comparison when Adriana says: "You're such a good father. I wish my dad had been like that." Indeed, Tony is evidently a much better father than his own, though his mother Livia is fond of declaring her husband Johnny to have been "a saint" while venting about the inadequacies of her son. (More on this in a moment.) Whether or not we think of Meadow and A. J. as innocent, they are certainly not moral monsters, and they love their father. Many other characters describe Tony as a good dad, husband, brother, and friend.

general reflections on the series' genre bending, see Polan, *"The Sopranos,"* 40–44, 108; and Carroll, "Sympathy for the Devil," 121.

[37] Gini, "Bada-Being and Nothingness," 14.

[38] Willis, "Our Mobsters, Ourselves," 3.

[39] Donatelli and Alward, "'I Dread You'?" 65.

[40] How much is the value of Tony's love undermined by his selfishness? By his (unconscious?) smokescreen of sentimentality? These are good questions, and the series asks them intelligently and consistently.

Besides loving and being loved, Tony has many charming traits. He is hilariously sarcastic. He says to A. J. when he is flirting with existentialism: "Even if God is dead, you're still gonna kiss his ass." He is a master of quick witticisms: "Well, sit down and dig into this medley of pastas that Janice whipped up." He is the *capo dei capi* of Soprano speak, that New Jersey–Italian patois "so compressed and inventive in its mix of tones and jargons that it sound[s] like a new dialect, a poetically charged speech welded out of obscenities and banalities, misconstrued catchphrases and newly minted messages from the unconscious."[41] He even has a goofy penchant for punning, which seems especially dissonant with his thoughtless malice. And, of course, he has the gift of gab.

Though this is less obvious, Tony's storytelling prowess is connected to his capacity for astute political vision. His cunning as an operator depends on discerning perception and a flair for imaginative narrativizing. *The Sopranos* presents this partly as hard-won wisdom—as when Tony seeks and internalizes advice from older mobsters such as Jackie Aprile and shrewd advisors such as Hesh Rabkin—and partly as an unteachable, intuitive grasp of subterranean realities, at certain points budding first in his active unconscious—as when the knowledge that Big Pussy is an informant comes to him in a dream. It is no coincidence, in short, that Tony becomes the boss of the DiMeo crime family. He is a savvy manager of his soldiers, a tough negotiator, and a preternaturally talented charmer.[42] He is also a tenacious, courageous warrior, regularly compared to a bull and an ox.

Additionally, Tony is something of a mental health awareness pioneer. He is suicidally depressed and suffers from panic attacks. In an obviously basic sense, the series is about his search for therapeutic help. He even brings his wife in for couples therapy. Though prudence restrains his ability to publicize this mental health journey, he does gain some moral credit for struggling to break free of his highly limiting milieu.

Facts about Tony's formative environment complicate our perspective on him at least as much as these personal charms. Consider Tony's toxic familial relations and especially the intimated details of his treatment as a child. His mother, Livia, is a classic victim turned villain, a "monster out of Balzac" whose misery expresses itself in joylessness, repression, nostalgic delusion, and

[41] O'Brien, "A Northern New Jersey of the Mind," 161.

[42] The contrast between Tony and his gang is emphasized when a trip to Italy sees Paulie mocked for classlessness, Christopher holed up in the hotel high on heroin, and a lone wolf Tony emerging triumphant from a sticky encounter with a female boss of the Neapolitan camorra. On Tony's uncharacteristic overcoming of sexual temptation in this scene, see Green, "'I Dunno about Morals, but I Do Got Rules,'" 67. Compare Polan, *"The Sopranos,"* 36.

misanthropy.[43] Whether or not Melfi is correct in diagnosing her with borderline personality disorder, she is certainly correct in urging Tony to reckon with the fact that his mother abused him emotionally in childhood ("I could stick this fork in your eye!"), continuously mocks and repudiates his desire for parental love ("Poor you!" is her favorite refrain), and vindictively orchestrates an attempt to assassinate him (Tony to Carmela: "What kind of person can I be, where his own mother wants him dead?").[44]

I am inclined to extend this analysis further. Tony's rage toward the mother who ignores his love—only acknowledged once he admits she wanted him dead—is a convenient cover for his more completely repressed rage toward his father, Giovanni "Johnny Boy" Soprano.[45] It was his father who denied him love most devastatingly, by serially abusing him, his siblings, and his mother, and by nudging him into the family business while Tony was still a child. In a scene whose psychic primacy is emphasized, Johnny praises his son for not expressing fear when he walks in on Johnny and his brother (Tony's uncle Corrado "Junior" Soprano) cutting off Satriale's pinkie finger as payment for a debt. This is the first step in Tony's rise to the top—and also, the series implies, in his descent to the bottom.

Melfi is perceptive in noticing from the start that Tony's desperate need for love is connected to his early family life and that the disappointment of this need leads to his own version of a split personality. He can seek love from Uncle Junior even after a feud over who should be head honcho eventuates in the attempted whacking. (Years later, Tony asks Junior, "Don't you love me?" after swearing over and over that "He's dead to me.") The germination of this hostility is one of the heights of narrative achievement in *The Sopranos*. Corrado has thrown away a joyful love affair that lasted for sixteen years. Why? Because his lover Roberta reveals to an acquaintance that Junior performs cunnilingus expertly, and the secret gets out. Mafioso misogyny being what it is, a predisposition to pleasing apparently renders even a boss unmasculine. "They think that if you suck pussy, you'll suck anything. It's a sign of weakness." Tony initially restrains himself, only mocking Junior in private, but eventually succumbs to slighting him publicly in retaliation after Junior needles him about going to therapy. Tony suspects he has made a mistake and expresses remorse in one of

[43] O'Brien, "A Northern New Jersey of the Mind," 162.

[44] "Poor you!" is repeated unknowingly by Gloria Trillo, one of Tony's many volatile mistresses, who is, as Melfi observes, much like his mother—to wit, a deeply damaged person who wants to die. Tony's wince when Gloria says this registers his uncomfortable recognition.

[45] Greene, "Is Tony Soprano Self-Blind?"

the most famous speeches in the show's history.[46] He is right: Junior sets up the hit attempt, one of the most crystallized expressions of the family dynamic just analyzed, both in the obvious sense that it is an instance of pathological intrafamilial violence and in the less obvious sense that being shot reinvigorates Tony (i.e., remasculinizes him) and temporarily quells his depression. Prozac and therapy cannot compare to the "kickstart" of warfare.[47]

The preceding reflections also begin to explain how the institution of masculinity is a powerfully corrosive force that stifles the moral development of all the male Sopranos. As Willis argues, Tony's gangsterism gives him a sense of power and control, excitement and action, and an outlet for his unacknowledged rage "without encroaching on his alter ego as benevolent husband and father."[48] But panic attacks and depression reveal the underlying conflict that alcoholism and sexual decadence can only intermittently conceal. *The Sopranos* is widely understood to be an investigation of the so-called "crisis of masculinity"—and the more general nostalgia for a lost postwar order characterized in part by its uncritical patriarchy.[49] In Tony's words, "Outside it may be the 1990s, but in this house it's 1954." Seeing how noxious but pervasive ideals of masculinity mediate Tony's psychological development rightly affects our judgments and sympathies.[50]

My point is not that a quick tongue or toxic family dynamic or the oppressiveness of masculinity *excuses* Tony's character. The point is that *The Sopranos* explores the moral development of evil across close to ninety hours of storytelling and that this exploration alters the complexion of our reactive attitudes. The parallel to Watson's story about Harris is undeniable, though I think Tony Soprano is far more interesting than Robert Harris.

Having explained why some sympathetic reactions to Tony are unobjectionable and desirable, I can now more usefully reconstruct and critique Eaton's

[46] "Uncle Junior and I, we had our problems with the business. But I never should have razzed him about eating pussy. This whole war could have been averted. Cunnilingus and psychiatry brought us to this!"

[47] Walker, "'Cunnilingus and Psychiatry Have Brought Us to This,'" 119.

[48] Willis, "Our Mobsters, Ourselves," 6.

[49] Lacey, "One for the Boys?"; Wolcott, "Bada Bing's Big Bang."

[50] In the next section, I will discuss another morally complicating feature of the narrative: its representation of masculinity in crisis gives female characters more agency than is traditional in cinematic depictions of the mafia (Donatelli and Alward, "'I Dread You'?"). Ironically, it can be argued that the work is a feminist improvement on the gangster film not just because it more seriously investigates pathological masculinity but also because it convicts some female characters of full-fledged complicity in organized crime. See Carmela's admission: "I have forsaken what is right for what is easy." And compare Valerie Palmer-Mehta's essay "Disciplining the Masculine" on Janice Soprano's feminine masquerade.

reasons for thinking that at least some of our reactions to Tony are morally suspicious.

Here is my understanding of the suggestions.[51] First, Eaton claims that some of us *love* or have great affection for Tony rather than merely sympathizing with him in the attenuated ways just canvassed.[52] Second, Eaton claims that we like Tony in part *because of his badness* and that we "take a strange satisfaction in his morally repugnant deeds."[53] Third, Eaton claims that we *take his side*, desiring that he commit more crimes, escape the police, and triumph over the forces of good.[54]

Were these claims correct, they would distinguish our reactions to Tony from our reactions to Harris. Suitably spelled out, they might undermine the interpretation I have sketched and, with it, the viability of my appeal to compartmentalized sympathy. Further, these claims depend on some stimulating moral psychological theses that are both controversial and underdiscussed. I will consider each of them in turn.

Can viewers love Tony Soprano? My preferred answer is that we can love him only in a special, nonliteral sense. We love Tony in the sense that we view him as an extraordinarily engaging fictional creation. He provides us with valuable aesthetic and moral experiences. We love watching him—that is, we love watching James Gandolfini play this role, and we love watching the show that revolves around him. We do not literally love Tony. According to almost all philosophers of love, love involves robust concern for the beloved and/or devotion to their good. It is hard to see how we could have this kind of relationship to a fictional character.[55]

We can set aside the question of whether to call this "love." The more immediate concern is that our reaction to Tony is supposed to involve morally bad forms of sympathy or identification—for instance, an affection that minimizes his flaws. An account is needed of why this would be. More specifically, we need an account of the kind of love that it is possible to direct at morally repugnant *fictional characters*. I have sketched one: this state is a special form of *appreciation*, that is, appreciation of the aesthetic experiences that their existence

[51] Some of this reasoning is stated in general characterizations of the rough hero, some in specific discussions of Tony.

[52] Eaton, "Robust Immoralism," 281.

[53] Eaton, "Robust Immoralism," 285, 287.

[54] Eaton, "Robust Immoralism," 285.

[55] I have defended a view about love's possible objects that is more permissive than the views of many philosophers (Shpall, "A Tripartite Theory of Love"). Nonetheless, I do not think my view can be extended to "love" for a fictional character, except in very special circumstances.

makes possible. This form of affectionate appreciation is probably distinctive to fiction. It is different in many respects from love for people who exist outside of fictions. Among other things, appreciating a character in this way is not so obviously connected to the danger of minimizing his flaws. Our appreciation might well involve being interested in keenly perceiving and evaluating such flaws. Fictional characters give us opportunities to appreciate in this way that real people usually do not.

Eaton's opening formulations are instructive in this connection. Consider the claim that we "miss" Tony. This claim is true on one natural understanding. The character was a stimulating, surprising narrative spectacle, and we were sad when the experience ended. It was also a delight to watch Gandolfini's incomparable performance, and it is easy to miss that too. Consider, by contrast, the claim that Tony "feels like an old friend." This is straightforwardly metaphorical. There are many ways in which Tony could not possibly feel like an old friend: friendship is necessarily reciprocal, we feel deeply alienated from our friends when they murder people, and so on. That we love or appreciate Tony in the special sense in which that is possible reveals no moral problem with our responses. The defender of Eaton's analysis must reject my suggestion about how to understand our love for Tony and propose another.[56]

Now consider Eaton's claim that we like Tony in part because of his badness. For me, this is the most compelling of Eaton's responses, though we disagree about its implications. Evaluating these issues requires a contentious foray into highly uncertain areas of philosophical psychology that philosophers writing on these issues have not yet pursued.

Attraction to badness in virtue of its (perceived) badness appears to be a real phenomenon. Here are some plausible examples from ordinary life. First, the disruptive humor of the class clown makes us laugh partly because the disruption is disrespectful. Second, for those of us seduced by drugs or other addictive substances or behaviors, such temptations may have a special magnetism precisely when we believe it is wrong to pursue them. Third, sexual fantasies frequently involve norm transgression, sometimes including the transgression of norms the fantasizers would never consider violating in real life.[57]

For me, even this preliminary catalog suffices to motivate the possibility of attraction to the bad. However, the examples also suggest, appropriately, that the psychological nature and normative status of this attraction are poorly

[56] See Eaton, "Robust Immoralism," 281. On the related metaphor of being friends with authors, compare Gaut, *Art, Emotion, and Ethics*, 109–12.

[57] See Aaron Smuts, "The Ethics of Singing Along," 125, for a related discussion and the claim that imaginative engagement with fiction involves fantasies that we do not want to be actualized.

understood. If this phenomenon is real, it is both fascinating and puzzling. Indeed, the topic has perplexed artists and philosophers for a long time.[58]

Let us assume that attraction to the bad occurs. If we observe it in ordinary life, how could it reveal something particularly significant about the responses we have to certain fictions? Suppose it is wrong to be attracted to Tony on account of his badness. Suppose for the time being that *The Sopranos* encourages this attraction (though I will dispute this in the next section). Still, this does not constitute an aesthetic achievement that increases the value of the series *qua* art. If attraction to badness on account of its badness is a common psychological phenomenon, then there is nothing aesthetically special about capitalizing on it. Genuine achievement requires success where success is not easy.[59]

So, the proponent of Eaton's argument must defend a more detailed account of the psychology of attraction to the bad, an account that vindicates the aesthetic achievement claim. It is not sufficient to say that overcoming our imaginative resistance is such an achievement, since attraction to the bad *always* involves overcoming resistance, even when it occurs in everyday contexts where any claim about artistry would be misplaced. It could be that artists who encourage our attraction to evil characters are simply tapping into abiding psychological dispositions that it is easy to activate. The argument's proponent must also defend an account of the norms on attraction to badness such that, for example, being attracted to Tony on account of his badness is morally problematic. In doing so, they will have to contend with various complications, for instance, the fact that sexual fantasies often seem to involve attraction to the bad and that it seems harsh to condemn these fantasies as morally problematic on these grounds alone.[60] Additionally, more needs to be said about whether and how we should differentiate attraction to the bad in fiction and fantasy from attraction to the bad in reality.

Finally, consider Eaton's claim that we take Tony's side, desiring that he commit more crimes, escape the police, and prevail over the forces of good.[61] To this claim there is an easy rejoinder. Tony is the heart of *The Sopranos*, which we enjoy watching tremendously. If Tony is killed, captured, or subdued, the

[58] Compare Poe, "The Imp of the Perverse"; Dostoevsky, *Notes from Underground*; Stocker, "Desiring the Bad"; and Velleman, "The Guise of the Good."

[59] To be clear, in this paragraph I am granting premises 1–5 for the sake of argument and denying premise 6 (the aesthetic achievement premise). In the next section, I explain how *The Sopranos* unwaveringly encourages *condemnation* of Tony's badness, even if it also encourages fascination with it.

[60] For an outstanding recent discussion of rape fantasies, see Fraser, "Rape Fantasies."

[61] See also Clavell-Vazquez, "Sugar and Spice, and Everything Nice." Clavell-Vazquez agrees that "appreciators are prescribed to *ally* with rough heroes."

series will end, or it will be radically transformed in ways that curtail our pleasure. We want him to commit more crimes in the sense that we want more crimes to happen in the world of Father Brown and that we want the bank heist in *Rififi* to continue to its successful consummation. If we take some satisfaction in his morally repugnant deeds, this is because those deeds are appreciated as apt continuations of an enjoyable yarn. Just as some aspects of our "love" for Tony are best understood as a distinctive form of aesthetic appreciation, our alliance with him is best understood as a desire that the filmmakers allow this appreciation to continue.[62]

What I emphatically deny is that most viewers take intrinsic enjoyment in the suffering of Tony's victims.[63] On the contrary, we are regularly appalled and disgusted by his crimes. I still feel a queasy repulsion whenever I recall the scene of Tony calling Adriana from a payphone to set up her horrifying murder by Silvio. Of course, moral crimes can be depicted with artistry, making their depiction enjoyable to watch even when we regard the depicted actions as abominable. I will later note that *The Sopranos* is often lauded for its grotesquely realistic and absurdly comic representations of violence, which resist the congenital cinematic temptation to romanticize violence for aesthetic ends.[64] I will also say more about how the series consciously departs from *The Godfather* and other mob cinema, including in its commitment to depicting the pathetic ugliness of evil. But these claims are unnecessary for the present argument.

I have explored several challenges to the idea that our sympathy for characters like Tony Soprano is morally problematic. I conclude this section by emphasizing how these reflections have also helped to locate some intriguing differences between sympathetic engagement with fiction and sympathetic engagement with reality.

We rarely desire to spend time with actual people we regard as despicable, though we may still like or admire some of their qualities. If Tony were really your neighbor, and after hearing his life story on National Public Radio you met him at a Parent-Teacher Association meeting, you might be curious and courageous enough to have a beer with him in a well-lighted place, but you probably would not invite him over for a family barbecue.[65] Your reasonable

[62] For an account of the underlying psychology of imaginative engagement with fiction and an application to "desiring the safety of Tony Soprano," see Doggett and Egan, "How We Feel about Terrible, Non-Existent Mafiosi," 290.

[63] For some provocative claims about whether artworks in general invite us to enjoy suffering, see Smuts, "The Ethics of Singing Along," 127.

[64] Polan, *"The Sopranos,"* 25–31.

[65] Carroll, "Sympathy for the Devil"; and Harold, "A Moral Never-Never Land."

hesitation would be grounded in distrust, fear, and moral condemnation—as it is for some fictional characters in *The Sopranos* who avoid Tony like the plague.

But *we* need not distrust or fear the fictional Tony Soprano, because he does not exist.[66] It would be unsurprising if this made it easier to like or admire him than it would be to like or admire a real-life Tony counterpart. A separate though related point is that the appropriateness conditions governing reactions to a fictional character with grievous moral flaws are surely different from those governing reactions to a real person with the same flaws. It is not just easier to like and root for the fictional Tony. It is less objectionable—if it is objectionable at all.[67]

What is more, there is an argument for thinking that most of us are too judgmental and hard-hearted in personal relations and that having more fictional sympathy, even for bad characters, could be a kind of virtuous training. I am not sure whether this is true. It does help highlight the range of concerns one might have with the claim that sympathy, affection, admiration, and other positive reactions to bad fictional characters are morally bad.

I have given my reasons for rejecting Eaton's second premise—that it is morally bad to sympathize with an irredeemably vicious character—when it is applied to Tony Soprano (assuming for the sake of argument that he is irredeemably vicious). As before, I acknowledge that a full analysis of the force of Eaton's argument needs to consider other cases. I doubt the essential points made here depend on idiosyncratic features of *The Sopranos*, but evaluation of this suspicion must be left to the enterprising reader.

4. THE CONTENT OF ARTISTIC PRESCRIPTIONS

I will now complete my argument by showing how difficult it is to make good on the claim of *immoral* artistry, that is, the claim that artworks—and especially great artworks such as *The Sopranos*—prescribe immoral responses in a way that contributes to their artistic achievement. I do not deny that some artworks prescribe some immoral responses. I do deny Eaton's claim that such

[66] Walton, "Fearing Fictions."

[67] Kieran, "Art, Morality, and Ethics," 135. This set of observations also puts pressure on Eaton's contention (premise 6) that rough-hero works are special aesthetic achievements. The special achievement is supposed to be a distinctive overcoming of imaginative resistance. But it may be that our interest in and sympathy for bad fictional characters has little resistance to overcome, because we recognize that our interest and sympathy is, on account of being directed at a fiction, unproblematic.

prescriptions are endemic to the works she identifies. It convicts too many artworks of too much immorality, and it does so in the wrong places.[68]

I will defend my skepticism by defending *The Sopranos* from the charge of immorality, articulating my own view about its most important prescriptions and their moral status. The interpretation I have offered already shows why we must ask more precise questions than "Does the series encourage us to sympathize with Tony?" The series encourages various forms of sympathy, condemnation, ambivalence, and many other attitudes and emotions.[69] I will develop my interpretation by defending three main theses about what reactions the filmmakers appear to be encouraging.

First, to the qualified extent that we are encouraged to identify with Tony, we are also encouraged to realize that our susceptibility to him is striking and problematic. In other words, we are encouraged to entertain an indictment of our dispositions and the (American) culture that has shaped them. *The Sopranos* is a sharp, absurdist critique of the moral-psychological foundations of American capitalism. We see ourselves not in Tony's murderousness but in his atavistic consumerism. More interestingly, we see our own perhaps inchoate misgivings dramatized as Tony vaguely grasps the tension between his unrestrained individualism in business and his valorization of family and community.[70]

At one point, Tony's mistress Gloria says: "You really are in love with yourself... you deprive yourself of nothing." This is only true in the sense that he takes whatever he wants, often prompting Carmela and Melfi to analogize him to a child—which highlights, not incidentally, the important connections between Tony and his father, who is almost always referred to as Johnny Boy. Properly understood, Tony's compulsive eating, drinking, and fucking are not

[68] For discussion of the distinction between an artwork that presents a perspective and one that expresses a commitment to it, see Giovanelli, "Ethical Criticism in Perspective."

[69] Compare the idea that *Paradise Lost* "prescribes our wonder, reverential admiration, and respect for the grand but evil being" (Eaton, "Robust Immoralism," 284). See McGregor, "A Critique of the Value Interaction Debate," in which he claims that Eaton oversimplifies both habitual reactions to Milton's Satan and the reactions prescribed by the work (449–66). Likewise, are we really lured into condoning the pedophilia of Humbert Humbert in *Lolita* (Eaton, "Robust Immoralism," 284)? For persuasive arguments against this, see Gaut, *Art, Emotion, and Ethics*, 194–202; and McGregor, "A Critique of the Value Interaction Debate," 457–58.

[70] The theme of callow consumerism is emphasized in numerous ways. Paulie's appalling crimes fund a life of watching mindless infomercials. Even Neapolitan mobsters delight in the ability to purchase Mont Blanc pens cheaply in New York because of the weak American dollar. The representation of gangsters as superficial, social-climbing bores departs markedly from cinematic traditions of representing them as rugged individuals or bastions of family values.

really expressions of self-love but of repressed despair.[71] Melfi suggests that "many Americans" identify with this brand of American malaise, with Tony's sense that he "came in at the end, that the best is over."[72] She is right. Viewers feel attached to this sentiment not because they lament the decline of the mafia but because they recognize Tony is really talking about the (perceived) decline of traditional sources of meaning: family, honorable and fair work, community, faith. They may also recognize that *The Sopranos* offers a diagnosis of the roots of these social changes—in secular capitalist individualism—that may emphasize our own collaboration.

Of course, some people are attracted to the nihilistic hedonism and apparently unaccountable patriarchy of Tony's world outside the home. That reaction raises interesting moral questions, and I will conclude this essay by briefly commenting on it. However, though Tony fantasizes that he can arrest social movement and sustain the old orders by force of will, this facade of control is represented as a hopeless death rattle. To the extent that *The Sopranos* encourages us to identify with Tony, it does so to make us recoil from our own problematic tendencies toward nostalgia and narcissism, and to laugh at the craven spectacle of hegemonic capitalism, whose apparently "law-abiding citizens" are often insider traders or tobacco executives.[73] The series does something morally serious in issuing such provocations.

Second, *The Sopranos* encourages viewers to reflect on American film's longstanding patterns of depicting and glorifying violence and, more specifically, on the meaning of cinema's obsession with the mafia.[74] Melfi's son Jason says: "At this point in our cultural history, mob movies are classic American cinema, like westerns."[75] Characters such as Silvio and Tony affectionately mimic gangster films, model themselves on their archetypes, and explicitly associate the rugged stoicism of the gangster with a broader tradition of masculine iconography, epitomized by Tony's compulsive habit of lionizing Gary Cooper via

[71] As Carmela's covetousness is a mask for her unhappiness—and an Achilles' heel that Tony regularly exploits to win back her good graces.

[72] Hayward and Biro, "The Eighteenth Brumaire of Tony Soprano": "Tony's problems are symptomatic of a more generalized cultural anxiety, or a more widely felt insecurity generated by commodification and the decline of community" (211). And it is not just Americans who so identify. Lacey, "One for the Boys?": "Tony Soprano functions as a cipher for the lived contradictions of the British middle-aged, or middle-aging, middle management lifestyle but with the escapist fantasies of Mafia masculinity" (100).

[73] Green, "'I Dunno about Morals, but I Do Got Rules,'" 61.

[74] Symonds, "Show Business or Dirty Business?"

[75] For more on how *The Sopranos* engages with the figure of the cowboy, see Polan, *"The Sopranos,"* 103–4; and Gini, "Bada-Being and Nothingness," 9.

the phrase "the strong, silent type."[76] Christopher does them one better by producing *Cleaver*, a ridiculous shlock-mafia-horror flick that premieres to a packed house of tickled mob families. Many critics have analyzed the aims of this intertextual play and the dozens of references in the series to films such as *Public Enemy* and especially *The Godfather* trilogy.[77] I will briefly sketch a core departure from tradition that supports the view that *The Sopranos* encourages a distinctive perspective on violent representation.

It is sometimes claimed that *The Godfather* anticipated the conservative "family values" revival of Reagan-era America.[78] The stately Corleone family and the Don who takes justice into his own hands when the state fails to live up to its promises expressed a germinating reactionary response to widespread feelings of social disintegration prompted, supposedly, by the 1960s counterculture and the upheavals of the civil rights and antiwar movements.[79] This context helps explain the Soprano crew's mythologizing. *The Godfather* presents at least a chosen few mobsters as refined and worldly-wise guardians of the family, justice, and Italian American identity.[80]

It also helps to contextualize the self-conscious exploration of violence in *The Sopranos* within the broader history of the gangster film. The domesticated gangster inhabits a narrative still characterized by outsized misogyny but in which women are now accorded "equal dramatic weight."[81] For example, Lorraine Bracco's casting as Jennifer Melfi functions to recall her role as Karen Hill in *Goodfellas*, which was itself "a breakthrough gangster film for the female

[76] The careful viewer sees cracks in this story. Compare James Harold on Tony's viewing of *Public Enemy*, his favorite film, after his mother's death, which prompts him to imagine having a loving mother, smile, and then cry ("A Moral Never-Never Land," 140); and Christopher Kocela on Tony's admission that he resents Dr. Melfi because the therapeutic relationship makes him feel like a pussy ("From Columbus to Gary Cooper," 106).

[77] Pattie, "Mobbed Up."

[78] Biskind, *Easy Riders, Raging Bulls*, 164.

[79] This is not to say that the filmmakers endorsed this reactionary perspective. Francis Ford Coppola thought of the film as an anti-capitalist allegory, though he also conceded that it projects an idealized vision of the mob (Cowie, *The Godfather Book*, 66).

[80] According to David Pattie, this also explains why Scorsese's mob films are not objects of adoration, even though they are evidently familiar with them ("Mobbed Up," 143). These films are "despairing, blackly humorous tales of a mob in decline," which too clearly express Tony's anxieties about the family business, the family, and society. The idealized mob of *The Godfather* and the America of the 1950s are—like the idealized version of Johnny Boy Soprano and the New Jersey mob of the 1970s—fictional Gary Coopers at the heart of Tony's self-constitution.

[81] O'Brien, "A Northern New Jersey of the Mind," 168.

narrator."[82] In *The Sopranos*, Melfi's repudiation of violence accelerates this humanizing critique of the genre. Besides foregrounding the "feminine," emotionally engaged, peaceful practice of psychotherapy in its very first scene, and besides sympathetically presenting Melfi's absorbed devotion to her patients, the series explores Melfi's own successful struggle to uphold her principles even when it is most tempting to violate them. After suffering a horrifying rape, Melfi refuses to reveal details about her injuries, evading Tony's prying even though police misconduct has led to a perversion of justice and she has confirmed the identity of the perpetrator. She understandably fantasizes about the retribution Tony would eagerly enact on her behalf. She openly explores her rage and vindictive desire with her own therapist. Nonetheless, this flirtation with temptation all the more convincingly frames her courage in resisting it and effectively reprimands those audience members who want her to acquiesce to Tony's corrupt system of justice.[83] The series prescribes respect and sympathy for this struggle and Melfi's strength of character. I will extend this crucial point momentarily.

More generally, *The Sopranos* centers the perspectives of female victims of patriarchal control in the form of central and not so central characters as different as Adriana and Tracee and Rosalie Aprile, characters who are to varying degrees complicit in the wrongdoing of those who oppress them.[84] And it depicts violent acts by employing a destabilizing mixture of tones, eschewing the aesthetic trappings that are often used to sanitize or even beautify them.[85] These are conscious, enduring departures from the implicit prescriptions surrounding depictions of violence in many classic works of American film and television in the gangster and western genres and beyond.

[82] Akass and McCabe, "Beyond the Bada Bing!" 148. Similarly, Suzanne Shepherd plays the mother of Karen Hill (Lorraine Bracco) in *Goodfellas* as well as Carmela's mother in *The Sopranos*. See Plourde, "Eve of Destruction."

[83] Baldanzi, "Bloodlust for the Common Man."

[84] It should be noted that Carmela is arguably the show's greatest character and, as I have suggested, a highly ambiguous figure in the context of this critique. She recognizes Tony's flaws, rebukes him, and often outwits him. She is an engaged mother, a caring friend, and a charitable figure in the community. But she is living high on blood money and she knows it—though she is able to maintain certain delicate fictions about whose blood is on her hands. This is put to her once with harsh directness by a psychoanalyst named Dr. Krakower. She weeps for a night, extracts $50,000 from Tony to donate to Columbia University, and takes refuge in the Catholic assurance that divorce is out of the question. For discussion, see McCabe and Akass, "What Has Carmela Ever Done for Feminism?" 47.

[85] On the grotesque, horrifying, and absurdly comic depictions of violence, which the creators employ to stimulate questions about our responses to violence in film and television, see Polan, *"The Sopranos,"* 25–31.

Third, Melfi is the moral center of the series. Notwithstanding her flaws, she is worthy of our sympathy and admiration. This identification colors all our reactions to the work. No plausible interpretation of the series and its moral prescriptions can ignore this pivotal fact.

Some critics make much of Melfi's fascination with Tony, which does sometimes veer into the prurient.[86] They may conclude partly on this basis that *The Sopranos* has no clear moral perspective and engages instead in a satirical postmodern sendup of bourgeois moralism.[87] I am not convinced. Questions about Melfi's professional obligations—whether she should have treated Tony in the first place, when she should have stopped—are left dangling, appropriately. Questions about her voyeuristic interest are emphasized throughout, with her therapist accusing her of seeking a "vicarious thrill." This set of questions is framed so starkly because Melfi's fascination mirrors our own.[88] But these observations do little to motivate any thoroughgoing skepticism about Melfi's character or her central role in articulating the ethical identity of *The Sopranos*.

Consider an interpretive puzzle at the heart of the work's conception. The puzzle concerns Melfi's faith in therapy. More specifically, it concerns her conviction that therapy has transformative potential even for someone as vicious as Tony.[89] Some characters believe this faith is naïve. Her ex-husband criticizes her (and her profession) of a "cheesy moral relativism" in the face of evil. Others suspect Melfi is just one more victim of Tony's expert manipulations.[90] At the very least, we are encouraged to wonder if her optimism owes more to motivated reasoning than evidence. Tony himself incessantly maligns the therapeutic process, lampoons Melfi's urbanity, and insultingly compares her to a useless con artist.

Yet Tony keeps coming to Melfi's office, paying what he claims to be extortionate rates for the privilege. Why? Because he knows Dr. Melfi is no dummy and no pushover. She is certainly no moral relativist. She condemns Tony more audaciously and insightfully than any other character, attacking his

[86] Polan, *"The Sopranos,"* 84, 121–22, 131.

[87] Polan, *"The Sopranos,"* 119.

[88] She even rightly chastises Dr. Eliot Kupferberg, *her own therapist*, who has often urged her to stop treating Tony, for returning to the topic of Tony in a session when she is speaking about unrelated things. The implication is that we are all in glass houses and should be careful about throwing stones. See Schulman, "An American Existentialism," 24. Schulman claims that Melfi is a "stand-in for the project of the show itself."

[89] Baldanzi, "Bloodlust for the Common Man," 89.

[90] Schulman, "An American Existentialism": "Psychoanalysis does not help Tony"; more generally, "no one on the show changes much" (34).

tribalism and his hypocrisy during many episodes of eloquent exchange.[91] She sees through Tony's defenses, urging him to repudiate his "high sentimentality mode" and to "own [his] feelings." Her brand of toughness is more convincing than his, at least to this viewer. She demonstrates astonishing poise and fortitude in dealing with an impossible patient—which Tony recognizes and respects to the point of thanking her sincerely for saving his life. And her brand of compassion is also compelling. She feels compassion for Tony not (primarily) because she is inappropriately fascinated by criminality or transfixed by his charms but because she understands him and his pain better than anyone else, even better than the people who know him best. This compassion is continuous with her compassion for her other deeply troubled and troubling patients, which is displayed vividly in the confrontation with Tony after Gloria's death. Finally, she is animated by an inspiring, philosophical-religious hope in the human potential to grow, particularly as a result of therapeutic intervention. She says that the process of talk therapy is "like giving birth." Revealingly, Tony accepts some aspects of this characterization while proposing an alternative metaphor: it is "more like taking a shit." In this context, that seems like a resounding endorsement.

Melfi's faith in self-examination and conversation may be overblown, in short, but it is also a beacon of sincerity and generous human feeling in a narrative world teeming with hypocrisy and egoism. Whether Tony does grow as a result of therapy is a more difficult issue. I am tempted to say that he improves in some modest ways and that Melfi deserves the credit.[92] This is an unpopular view, which Melfi herself apparently repudiates as the story ends. Having finally terminated their sessions, she says that psychopaths like Tony "sharpen their skills as con men on their therapists." That does not settle matters, but I need not defend any characterization of Tony's moral trajectory. The important thing is that Melfi is the perspectival core of *The Sopranos,* the character with whom we are encouraged to identify most. We are prescribed to use her thoughts and feelings as (fallible) guides to the moral realities of the fiction—and as checks on our own voyeuristic impulses.

[91] Harold, "A Moral Never-Never Land"; and Plourde, "Eve of Destruction."

[92] One example: Melfi appears to succeed in softening Tony's homophobia, *almost* to the point of getting him to pardon Vito Spatafore and welcome him back into the fold. It is the powerful homophobic hatred of others—his captains, the Leotardo family, even Carmela—that eventually forces Tony's hand. Another example: Tony's impulse control does seem to improve, if slightly. It is almost unbelievable that he refuses to sleep with Julianna, but he does appear to walk away from the encounter because of his desire to honor Carmela's devotion to him. Larger and more difficult examples concern the evolutions in Tony's relationships with characters such as A.J. and Janice.

I have now provided an interpretation of *The Sopranos* and an analysis of some of its structuring prescriptions that show the series to be pursuing a variety of artistic strategies that are *prima facie* morally praiseworthy. How might the immoralist respond? I will close by considering two objections. I hope this demonstrates how my analysis of the series facilitates useful responses that help clarify the state of play in value interaction debates.

Noel Carroll and others emphasize that there is distinctive value in artworks that expose our own limitations as moral reasoners.[93] Eaton thinks this view commits us to inconsistency.[94] In order to claim that works such as *The Sopranos* serve as a cautionary warning about irrelevant moral static, it must be that they do in fact lead us to, for example, inappropriately minimize the moral failings of characters such as Tony, at least for a time. So, immoral prescription is an essential condition of these works' achievements.

The Sopranos is an interesting test case here. On the one hand, it is particularly adept at prompting viewers to reflect on how personal charm can distort patterns of affective response and the causally related processes of moral judgment. On the other hand, it never shrinks from presenting the disgusting, appalling brutality of evil. Some viewers may minimize Tony's awfulness because they are charmed by his jokes or stupefied by his magnetism. Yet that is their mistake, and we learn something from it.[95] Most viewers are not at all confused about how bad Tony is, as I argued at length in the previous section. That suggests the series is less invested in uncovering flaws in our moral responses than some have believed. I think it is much more invested in ironically distancing us from a relatively uncritical interest in cinema's mobsters, that is, from more traditional cinematic representations of violent men. In any event, even if the series does encourage local minimization of Tony's flaws, this does not suffice for immoral prescription, assuming that *The Sopranos* encourages it in order to facilitate valuable reflection.[96]

[93] Carroll, "Rough Heroes." On artworks that dramatize "how easily we can be moved to take up attitudes we would reject if we thought more carefully" and that show us "the manipulative power of rhetoric in general and of art in particular," see Gaut, *Art, Emotion, and Ethics,* 201. Compare George Wilson on *Letter from an Unknown Woman* (*Narration in Light,* ch. 6) and Gaut on *The Destructors* and *Lolita* (*Art, Emotion, and Ethics,* 192–202). For a recent discussion of "seductive artworks," see Stear, "The Paradox of Seductive Artworks."

[94] Eaton, "Reply to Carroll," 376.

[95] On bad fans, see Nussbaum, "The Great Divide." On the cult of *Scarface* worship and how it subverts the intentions of the film's creators, compare Smuts, "The Ethics of Singing Along."

[96] Some claim that even if the higher-level artistic aim is morally good, and even if the lower-level ambition to stimulate morally problematic judgments is instrumental to this aim, the lower-level aim suffices for immoral artistry (Kieran, "Art, Morality, and Ethics,"

Eaton is also skeptical about the claim that we do successfully compartmentalize our reactions to characters such as Tony. This skepticism is grounded in her account of what rough hero works characteristically prescribe. Here is an interesting passage:

> It is, on my account, a paradigmatic feature of [rough hero works] that they deliberately make it nearly impossible for us to resolve the conflict between our approval and disapproval by neatly cordoning off the deplorable from the admirable … good instances of the Rough Hero solicit a powerful cocktail of pro attitudes directed at a complex multiplicity of intertwined traits, and this serves to land the audience in a state of deep ambivalence and moral confusion: we approve of something that we also condemn and are kept from settling on any consistent position.[97]

How exactly might *The Sopranos*, or any artwork, encourage *morally problematic inconsistency* in moral judgment about its characters? The key hypothesis is that there is "something" we are prompted to both approve of and condemn, which renders ambivalence deep and irresolvable and compartmentalization impossible. This something is Tony Soprano himself or a set of his traits.[98]

It is my contention that detailed interpretation undermines this claim about the structure and moral status of our ambivalence by isolating many of the separably evaluable components of his character as well as potentially mitigating facts about his history and milieu. Moreover, *The Sopranos* is a great work of art partly because it prescribes this complex yet explicable and consistent suite of reactions. Consider this passage from my favorite essay on the series:

> By the time we got to the end we had seen a thousand Tonys—sheepish, serpentine, commanding, calculating, lecherous, self-pitying, savagely sarcastic, tenderly paternal, fatuously self-pleased, teary-eyed over an old radio hit, racked by paranoid mistrust, exploding in feral rage—and seen one switch to another in an instant. Guileless self-revelation was not a possibility, least of all in a psychiatrist's office. *He had so many of him to choose from.*[99]

Such clarity of analysis does not suggest moral confusion. Tony charms us, beguiles us, repulses us, and we happily submit to it all. We are not often

139). I do not see why. For a useful recent discussion, see Stear, "The Paradox of Seductive Artworks," 478.

[97] Eaton, "Reply to Carroll," 378.

[98] Eaton, "Reply to Carroll": "We are moved to *simultaneously approve and disapprove of the same character* yet are offered nothing to resolve the conflict" (379, emphasis added).

[99] O'Brien, "A Northern New Jersey of the Mind," 162 (emphasis added).

confused about whether and when to condemn him or about which of his traits are good (tenderly paternal) and which bad (feral rage). But we are often entranced by this artwork's capacity to give even evil characters the sorts of traits and histories that inspire justified admiration and sympathy.

The Sopranos shows how difficult it is to make good on the claim of immoral artistry. If the series encourages immoral reactions, they are not the ones Eaton suggests. Here are two better candidates. First, it may encourage the stereotyping of Italian Americans. I will not defend a view about this topic. I will note that the series regularly engages in metafictional play aimed at drawing attention to both the ills of prejudice and the often sanctimonious, hypocritical follies of ethnic pride. I also note that David Chase, the show's creator, has addressed criticism from Italian Americans forcefully and with intelligence.[100] Suppose for the sake of argument that the series does encourage pernicious anti–Italian American stereotypes. It would be an uphill battle to make this moral flaw into an artistic virtue, even if those stereotypes are necessary for the work's existence. As I argued at the beginning of this paper, racist punchlines do not on their own constitute valuable artistic strategies.

Second, *The Sopranos* may encourage viewers to identify with a misogynistic ideal of manhood and, particularly, with a problematic ideal of male sexual conquest and what facilitates it. For example, it may do this by too uncritically bringing attractive women into the sexual-romantic orbit of Tony and other sociopathic criminals.[101] It is tempting to note in reply that Tony's affairs are much less satisfying than he hopes (indeed, some are positively traumatizing); that the problems with gendered beauty standards in film and television are general ones, in no way specific to this artwork; and that power, mysteriousness, and even dangerousness do in fact contribute to sexual attraction, presumably in gendered ways. But I can see an argument that *The Sopranos* overplays its hand here. If it does encourage a crassly misogynistic vision of male achievement, then I think this is a dual defect, an unfortunate deviation from the moral

[100] See Peter Bogdanovich's terrific interview with Chase, especially the discussion beginning at 1:05 (Bogdanovich, "Exclusive Video Interview with Sopranos Creator David Chase").

[101] In a sharp bit of intertextual commentary, this view—that mob films not only display a preexisting masculine fantasy of sexual promiscuity and objectification but actively create and perpetuate it—is forcefully articulated in *The White Lotus* (season 2), a recent HBO series, by Albee Di Grasso, the young adult son of the sex-addicted Dominic Di Grasso. Dominic is played by Michael Imperioli, who everybody associates with Christopher Moltisanti. Though Albee is there critiquing his father's and grandfather's adoration of *The Godfather*, the audience knows that in some sense he must also be commenting on *The Sopranos*.

and aesthetic standards of an otherwise extraordinary contribution to American culture.

University of Sydney
sam.shpall@sydney.edu.au

REFERENCES

Akass, Kim, and Janet McCabe. "Beyond the Bada Bing! Negotiating Female Authority in *The Sopranos*." In Lavery, *This Thing of Ours*, 146–61.

Baldanzi, Jessica. "Bloodlust for the Common Man: *The Sopranos* Confronts Its Volatile American Audience." In Lavery, *Reading "The Sopranos,"* 79–89.

Biskind, Peter. *Easy Riders, Raging Bulls: How the Sex-Drugs-and-Rock-'n'-Roll Generation Saved Hollywood*. New York: Simon and Schuster, 1999.

Bogdanovich, Peter. "Exclusive Video Interview with *Sopranos* Creator David Chase." *The Sopranos: The Complete First Season*. DVD. HBO, 2000.

Carroll, Noel. "Rough Heroes: A Response to A. W. Eaton." *Journal of Aesthetics and Art Criticism* 71, no. 4 (November 2013): 371–76.

———. "Sympathy for the Devil." In Greene and Vernezze, *"The Sopranos" and Philosophy*, 121–36.

Clavell-Vazquez, Adriana. "Sugar and Spice, and Everything Nice: What Rough Heroines Tell Us about Imaginative Resistance." *Journal of Aesthetics and Art Criticism* 76, no. 2 (April 2018): 201–12.

Cormier, Harvey. "Bringing Omar Back to Life." *Journal of Speculative Philosophy* 22, no. 3 (2008): 205–13.

Corwin, Miles. "Icy Killer's Life Steeped in Violence." *Los Angeles Times*, May 16, 1982.

Cowie, Peter. *The Godfather Book*. Boston: Faber and Faber, 1997.

Doggett, Tyler, and Andy Egan. "How We Feel about Terrible, Non-Existent Mafiosi." *Philosophy and Phenomenological Research* 84, no. 2 (March 2011): 277–306.

Donatelli, Cindy, and Sharon Alward. "'I Dread You'? Married to the Mob in *The Godfather*, *Goodfellas*, and *The Sopranos*." In Lavery, *This Thing of Ours*, 60–71.

Dostoevsky, Fyodor. *Notes from Underground*. Translated by Richard Pevear and Larissa Volokhonsky. New York: Vintage Classics, 1993.

Eaton, A. W. "Reply to Carroll: The Artistic Value of a Particular Kind of Moral Flaw." *Journal of Aesthetics and Art Criticism* 71, no. 4 (November 2013): 376–80.

———. "Robust Immoralism." *Journal of Aesthetics and Art Criticism* 70, no. 3 (August 2012): 281–92.

Fraser, Rachel. "Rape Fantasies." *The Point,* January 28, 2020. https://thepointmag.com/criticism/rape-fantasies/.

Gaut, Berys. *Art, Emotion, and Ethics*. Oxford: Oxford University Press, 2007.

——— . "The Ethical Criticism of Art." In *Aesthetics and Ethics: Essays at the Intersection,* edited by Jerrold Levinson, 182–203. Cambridge: Cambridge University Press, 1998.

Gendler, Tamar. "The Puzzle of Imaginative Resistance." *Journal of Philosophy* 97, no. 2 (February 2000): 55–81.

Gini, Al. "Bada-Being and Nothingness: Murderous Melodrama or Morality Play?" In Greene and Vernezze, *"The Sopranos" and Philosophy,* 7–14.

Giovanelli, Alessandro. "Ethical Criticism in Perspective: A Defense of Radical Moralism." *Journal of Aesthetics and Art Criticism* 71 (Fall 2013): 335–48.

Green, Ronald M. "'I Dunno about Morals, but I Do Got Rules': Tony Soprano as Ethical Manager." In Greene and Vernezze, *"The Sopranos" and Philosophy,* 59–71.

Greene, Richard. "Is Tony Soprano Self-Blind?" In Greene and Vernezze, *"The Sopranos" and Philosophy,* 171–81.

Greene, Richard, and Peter Vernezze, eds. *"The Sopranos" and Philosophy: I Kill Therefore I Am*. Chicago: Open Court, 2004.

Harold, James. "A Moral Never-Never Land: Identifying with Tony Soprano." In Greene and Vernezze, *"The Sopranos" and Philosophy,* 137–46.

Hayward, Steven, and Andrew Biro. "The Eighteenth Brumaire of Tony Soprano." In Lavery, *This Thing of Ours,* 203–14.

Holden, Stephen. Introduction to *The New York Times on "The Sopranos."* New York: ibooks, 2000.

Jacobson, Daniel. "Ethical Criticism and the Vice of Moderation." In *Contemporary Debates in Aesthetics and the Philosophy of Art,* edited by Matthew Kieran, 342–55. Chichester: Blackwell, 2005.

———. "In Praise of Immoral Art." *Philosophical Topics* 25, no. 1 (Spring 1997): 155–99.

John, Eileen. "Artistic Value and Opportunistic Moralism." In *Contemporary Debates in Aesthetics and the Philosophy of Art,* edited by Matthew Kieran, 332–41. Chichester: Blackwell, 2006.

Kieran, Matthew. "Art, Imagination, and the Cultivation of Morals." *Journal of Aesthetics and Art Criticism* 54, no. 4 (Autumn 1996): 337–51.

———. "Art, Morality, and Ethics: On the (Im)moral Character of Artworks and Inter-Relations to Artistic Value." *Philosophy Compass* 1, no. 2 (March 2006): 129–43.

———. “Forbidden Knowledge: The Challenge of Immoralism.” In *Art and Morality,* edited by Jose Luis Bermudez and Sebastian Gardner, 56–73. London: Routledge, 2003.

Kocela, Christopher. “From Columbus to Gary Cooper: Mourning the Lost White Father in *The Sopranos*.” In Lavery, *Reading “The Sopranos,”* 104–17.

Lacey, Joanne. “One for the Boys? *The Sopranos* and Its Male British Audience.” In Lavery, *This Thing of Ours,* 95–108.

Lavery, David. “‘Coming Heavy’: The Significance of *The Sopranos*.” In Lavery, *This Thing of Ours,* xi–xviii.

———, ed. *Reading “The Sopranos”: Hit TV from HBO*. London: I. B. Tauris, 2006.

———, ed. *This Thing of Ours: Investigating “The Sopranos.”* London: Wallflower Press, 2002

Li, Zhen. “Immorality and Transgressive Art: An Argument for Immoralism in the Philosophy of Art.” *Philosophical Quarterly* 71, no. 3 (July 2021): 481–501.

McCabe, Janet, and Kim Akass. “What Has Carmela Ever Done for Feminism? Carmela Soprano and the Post-Feminist Dilemma.” In Lavery, *Reading “The Sopranos,”* 39–55.

McGregor, Rafe. “A Critique of the Value Interaction Debate.” *British Journal of Aesthetics* 54, no. 4 (October 2014): 449–66.

Mermelstein, Jeff . *#nyc*. London: MACK, 2020.

Nannicelli, Ted. “Moderate Comic Immoralism and the Genetic Approach to the Ethical Criticism of Art.” *Journal of Aesthetics and Art Criticism* 72, no. 2 (October 2014): 169–79.

Nussbaum, Emily. “The Great Divide.” *New Yorker,* March 31, 2013.

O’Brien, Geoffrey. “A Northern New Jersey of the Mind.” In *Stolen Glimpses, Captive Shadows*. Berkeley, CA: Counterpoint, 2013.

Palmer-Mehta, Valerie. “Disciplining the Masculine: The Disruptive Power of Janice Soprano.” In Lavery, *Reading “The Sopranos,”* 56–68.

Paris, Panos. “The ‘Moralism’ in Immoralism: A Critique of Immoralism in Aesthetics.” *British Journal of Aesthetics* 59, no. 1 (January 2019): 13–33.

Pattie, David. “Mobbed Up: *The Sopranos* and the Modern Gangster Film.” In Lavery, *This Thing of Ours,* 135–45.

Pevear, Richard. Foreword to *Notes from Underground,* by Fyodor Dostoevsky, translated by Richard Pevear and Larissa Volokhonsky, vii–xxiii. New York: Vintage Classics, 2003.

Plourde, Bruce. “Eve of Destruction: Dr. Melfi as Reader of *The Sopranos*.” In Lavery, *Reading “The Sopranos,”* 69–76.

Poe, Edgar Allen. “The Imp of the Perverse.” In *The Complete Stories,* 271–75. New York: Everyman’s Library, 1993.

Polan, Dana. *“The Sopranos.”* Durham, NC: Duke University Press, 2009.

Schulman, Alex. "The Sopranos: An American Existentialism?" *Cambridge Quarterly* 39, no. 1 (March 2010): 23–38.

Shelby, Tommie. "Justice, Deviance, and the Dark Ghetto." *Philosophy and Public Affairs* 35, no. 2 (Spring 2007): 126–60.

Shpall, Sam. "A Tripartite Theory of Love." *Journal of Ethics and Social Philosophy* 13, no. 2 (May 2018): 91–124.

Shuster, Martin. *New Television: The Aesthetics and Politics of a Genre*. Chicago: University of Chicago Press, 2017.

Smuts, Aaron. "The Ethics of Singing Along: The Case of 'Mind of a Lunatic.'" *Journal of Aesthetics and Art Criticism* 71, no. 1 (February 2013): 121–29.

Song, Moonyoung. "The Nature of the Interaction between Moral and Artistic Value." *Journal of Aesthetics and Art Criticism* 76, no. 3 (August 2018): 285–95.

Sontag, Susan. "Fascinating Fascism." In *A Susan Sontag Reader*, 305–25. New York: FSG, 1982.

Stear, Nils-Hennes. "Immoralism Is Obviously True: Towards Progress on the Ethical Question." *British Journal of Aesthetics* 62, no. 4 (October 2022): 615–32.

———. "The Paradox of Seductive Artworks." *Australasian Journal of Philosophy* 97, no. 3 (July 2019): 465–82.

Stecker, Robert. "Immoralism and the Anti-Theoretical View." *British Journal of Aesthetics* 48, no. 2 (April 2008): 145–61.

Stocker, Michael. "Desiring the Bad: An Essay in Moral Psychology." *Journal of Philosophy* 76, no. 12 (December 1979): 738–54.

Symonds, Gwyn. "Show Business or Dirty Business? The Theatrics of Mafia Narrative and Empathy for the Last Mob Boss Standing in *The Sopranos*." In Lavery, *Reading "The Sopranos,"* 127–37.

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

Walker, Joseph S. "'Cunnilingus and Psychiatry Have Brought Us to This': Livia and the Logic of False Hoods in the First Season of *The Sopranos*." In Lavery, *This Thing of Ours*, 109–21.

Walton, Kendall. "Fearing Fictions." *Journal of Philosophy* 75, no. 1 (January 1978): 5–27.

Watson, Gary. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Agency and Answerability*, 219–59. Oxford: Oxford University Press, 2004.

Willis, Ellen. "Our Mobsters, Ourselves." In Lavery, *This Thing of Ours*, 2–9.

Wilson, George. *Narration in Light*. Baltimore: Johns Hopkins Press, 1986.

Wolcott, James. "Bada Bing's Big Bang: How *The Sopranos* Defined White America's Cultural Shift." *Vanity Fair*, November 30, 2018.

Journal of Ethics and Social Philosophy
vol. 27, no. 1 · march 2024

https://doi.org/10.26556/jesp.v27i1.2888
© 2024 Author

WEAKNESS OF POLITICAL WILL

Camila Hernandez Flowerman

Nations often fail to act in accordance with their best interests. In most cases, there is no great mystery behind this—collective action is difficult and complicated, and there are many trade-offs and resource constraints that limit what kind of action is possible for governments and institutions to accomplish. In some cases, though, the failures are puzzling. In such cases, it appears *literally possible* for a government or other political institution to take action, and yet it fails to do so, even when such actions would seem to be desirable and perhaps even morally obligatory. Consider climate change: even as many nations suffer dire consequences in the form of increasingly intense weather-related natural disasters, many of those same nations repeatedly fail to commit themselves to policies or legislation that would meaningfully address the crisis. The puzzling feature of this kind of case is that the failure in question seems to be the result of a *motivational* or *volitional* limitation. Failures occur not because of any material or resource limitation, but because the governing collective cannot will itself to realize the set of actions that would bring about a desirable state of affairs.

Yet in spite of a well-developed philosophical literature on the concept of *akrasia* or weakness of will as it appears in individual agents, there exist relatively few accounts of the same phenomena in collective political agents such as governments.[1] And while claims about insufficient "political will" are often

[1] Notable exceptions include discussions of collective *akrasia* in Pettit, "Akrasia, Collective and Individual," 68–96; and List and Pettit, *Group Agency*, ch. 9. Interestingly, there has been some work on other kinds of motivational issues that inhibit our ability to address the climate crisis, for example in Michael Doan's work on complacency and motivational vice in the face of climate change. Doan argues that complacency is a form of "motivational inertia" where agents are "caught up in patterns of behavior that expressed settled expectations of self-sufficiency" ("Climate Change and Complacency," 11). Complacency and weakness of will are, by that definition, distinct concepts. But it seems plausible to me that they have a kind of "family resemblance" in that both are related to motivation and can do similar work of offering an explanatory analysis of failures to act in the case of climate change. In the political case, however, I think weakness of will looks like a more plausible explanation for *collective* failure to act, or inaction of group agents. Note also that some

used to explain a lack of action on a variety of political issues, the term itself is imprecise, and it is unclear what commentators actually mean when they employ the concept. In this paper, therefore, my aim is to provide a preliminary account of weakness of political will (political *akrasia*). In doing so, I will also articulate and argue for a more expansive concept of political will in order to better account for the many different points at which a political agent might suffer a breakdown of that will.

1. Group Agency, Political Collectives

Thus far I have not defined the term "political agent." You might wonder why we should think of groups like nations, political institutions, etc., as agents in their own right. We certainly talk about such groups as though they have agency, or at least the ability to act and bear responsibility. Even the term "political will" seems to imply that such political actors have "wills" that might behave in much the same way our individual wills do. But perhaps such talk is merely a manner of speech, and what we really mean to say about groups is that they are just amalgamations of the wills and actions of their constituent group members. In other words, group "agency" may just reduce fully to the combined agency of individual members of the group. If this is the case, then it would seem that group weakness of will might just be explainable by reference to weakness of will in the individual members that constitute the group. Those who are skeptical of talk about group agency or collective agents may therefore find the concept of weakness of political will hard to buy.

However, I think it is fairly clear that there *are* group agents in their own right, where this agency exists over and above, and is not reducible to, the agency of constituent group members, and, further, that such agents can experience a kind of phenomenon similar in type to that of weak-willed individuals. But more needs to be said about what exactly this agency consists in, and what kinds of political entities count as agents in their own right, in order to make sense of the term "political agent."

Consider the following distinction. Sometimes there are collections of people who happen to be together but are not unified in any sense. These are what Christian List and Philip Pettit refer to as "mere collections."[2] Other times there are what Virginia Held calls a "random collection of individuals," or collections of individuals who may appear unified in the sense that they act

consider *akrasia* and weakness of will to be distinct concepts. For simplicity's sake, I use the terms interchangeably.

[2] List and Pettit, *Group Agency*, 31.

together in some instance, but they are not a part of an identifiable group that could be considered a "group agent."[3]

One thing that distinguishes cases of group agency from these cases of mere small-scale shared agency among collections of individuals is the fact that these larger-group agents appear to persist through time even as the individual agents that constitute the group may change or even turn over completely. Even though all groups are made up of individual agents, groups such as companies, universities, committees, etc., seem to have a kind of "corporate identity" that transcends individual membership.[4] Consider, for example, the Supreme Court of the United States. The Supreme Court's membership has changed completely over the history of the United States, but the court's identity as the Supreme Court has remained the same. Though the existence of these groups may be dependent on the participation of individual members in the group, the group is not identical to those members. Just as a lump of clay "is not identical to the statue, but it is the material of which the statue is made," the current members of the Supreme Court "are not identical to the Supreme Court, but they are the people out of whom the Supreme Court is made."[5]

Another feature of these kinds of "corporate" agents that makes them distinct from smaller-scale cases of shared agency is that, though they may operate through their members, their operations appear to be distinct from the actions of any individual members. As Held explains, an organized group agent is distinguishable not just by "certain characteristics that delimit its membership from other persons, but especially by its possession of a decision method for acting."[6] Group agents of this kind might have certain goals or aims. Further, they may also, through various procedures, arrive at judgments about how best to achieve those aims. They may even have methods of reviewing or updating those aims, along with their judgments about how to achieve them.[7] All of this might be true without any individual constituent member or set of members sharing those aims or arriving at those judgments.

[3] Similar to Held, Larry May refers to these groups as "loosely structured groups." This can also be distinguished from groups that have what Scott Shapiro calls "massively shared agency," or those involving the agency of many participants but that cannot be understood as a "group agent" in their own right. See, May, "Collective Inaction and Shared Responsibility," 269; Shapiro, "Massively Shared Agency," 257–93; and Held, "Can a Random Collection of Individuals Be Morally Responsible?" 471.

[4] Pettit, "Responsibility Incorporated," 172.

[5] Epstein, *The Ant Trap*, 142.

[6] Held, "Can a Random Collection of Individuals Be Morally Responsible?" 471.

[7] Pettit, "Responsibility Incorporated," 172.

In order to better illustrate this, consider Philip Pettit's "discursive dilemma" as an example of a case where a group's attitudes come apart from the attitudes of individual constituent members. Pettit asks us to imagine a company that takes up the question of whether to use the money originally intended to be a pay raise to instead introduce a set of workplace safety measures against some danger. He also asks us to suppose that the employees are going to make the decision on the basis of three considerations: whether the danger is serious, whether the safety measure will be effective, and whether the pay sacrifice is bearable.[8] Employees *A, B,* and *C* each deliberate on the particular premises (serious danger, effective measure, bearable loss), and then vote for a particular conclusion on the basis of their deliberation on those premises. Only if the employees think the answer is yes to each premise will they reason in favor of the pay sacrifice. Table 1 is modeled off the one Pettit provides to show this.[9]

Table 1.

	Is there serious danger?	Will the safety measure be effective?	Would the pay sacrifice constitute a bearable loss?	*Should we accept the pay sacrifice?*
A	Yes	No	Yes	*No*
B	No	Yes	Yes	*No*
C	Yes	Yes	No	*No*

The group might deliberate through what Pettit calls a "conclusion-centered" option (which takes into consideration only each group member's final conclusion about the pay sacrifice), or through what he refers to as a "premise-centered" option (which takes into consideration their conclusions about each of the three premises). If they deliberate through the former, then the result will be against the pay sacrifice, as each individual member has concluded against the pay sacrifice. But if they deliberate instead through the latter, for example if there was a chairperson who took a vote on each of the premises, then the result might be in favor of the pay sacrifice, as there are more positive responses under each premise than there are negative.[10] If they arrive through their standard process of decision-making at the latter conclusion, then the group will decide in favor of the pay sacrifice in spite of the fact that no individual member of the group decided (on their own) in favor of the pay sacrifice. Thus the group

[8] Pettit notes that they might do this because of "prior resolution." Once the group sets out its rules for decision-making, in effect those rules *become* the group deliberative faculty.

[9] Pettit, "Groups with Minds of Their Own," 171.

[10] Pettit, "Groups with Minds of Their Own," 171.

deliberative processes can come apart from the individual deliberative conclusions of constituent group members.

Particularly when group deliberative processes are complicated, extended, and involve multiple factors, the deliberative faculty that results in action for the group can be distinct from the ones involved in the rational deliberative process for any particular individual, or even for all of the individuals involved in the group.

Groups that have these features—the ability to persist through time regardless of membership changes, and a decision-making procedure that produces attitudes or judgments not directly attributable to any particular member or set of members—can be understood as group agents in their own right.[11] As I use the term in this paper, a political agent must be an instance of this kind of group in order to be capable of acting, and therefore to be capable of suffering from weakness of will that prevents it from completing the action that accords with its better judgment.

But what sorts of groups in our actual world count as political agents? Some instances of political collectives are more clearly identified as group agents than others. For example, the Supreme Court is a political collective that persists through membership changes and whose attitudes are not directly reducible to the attitudes of each court member, given its formal decision-making procedures. The same is true for entities such as political parties, the United States Senate, the United States government as a whole, etc. What is more controversial is whether "the people" in aggregate, where that refers to the individual residents and constituent members of a state, constitute a group agent. Furthermore, whether or not "the United States" itself can count as a group agent is not obvious, though we frequently talk about it as though it is.

For the purposes of simplicity, in this paper I will limit the term "political agent" to refer to just these collectives that have a readily identifiable decision-making procedure and member-independent identity.[12] When I refer to "the United States," then, this is really just a way of referring to its governing collective. For reasons that will become clear later in the next section, I think

[11] Given certain views about group agency, these two criteria may not be completely independent of one another. For example, Tracy Isaacs notes that "the more structure a collective has, the easier it is to dissociate its identity from any particular cohort of group members" (*Moral Responsibility in Collective Contexts*, 24).

[12] Isaacs differentiates between "organizations" on the one hand and "goal-oriented collectives" on the other. I think it is likely that most political agents will turn out to be organizations under this definition, but because both types of groups can be said to have a "collective intentional structure that gives rise to collective intention and action," it would also be fine for my argument if some turn out to be goal-oriented collectives, as opposed to organizations (*Moral Responsibility in Collective Contexts*, 27).

that our conception of the governing collective in this case will need to be broadened to include not just specific aspects of a government, but the system of governmental and institutional actions that ultimately affect the production of policy and legislation. For example, the Supreme Court is itself a political agent in virtue of it acting in the political sphere (through its role in protecting or overturning legislation based on legal criteria). It is also a subgroup (constituent member) of the larger political agent, the United States government. We can therefore analyze the ways in which the Supreme Court could itself be weak willed; alternatively, we could also analyze the ways in which (even through proper internal functioning) the Supreme Court might play a role in the United States being weak willed, given its status as an individual constituent member of the larger group agent.[13] The general idea is that a political collective is a group agent with a member-independent identity that acts in the political sphere, through a decision-making procedure, often (but not always) to produce, evaluate, or otherwise impact policy and legislation.[14]

2. Clarifying the Concepts

To begin with, I want to distinguish between two ways of talking about the will. The distinction is basically this: sometimes we talk about "the will" as a kind of general power or capacity, while other times we talk about "a will" to do something in particular. These two understandings are not mutually exclusive, provided that "the will" and "a will" are meant to apply to different concepts. An

[13] Similarly, on a narrow view of political parties, where "the Republican Party" is understood as something like the Republican National Committee, political parties will also count as a political agent. The Republican Party has a member-independent identity (it persists through time regardless of changes in membership), and it has both formal and informal rules for decision-making within the political sphere. For example, it has formal decision-making procedures for nominations to political positions at various levels of government, but it also has informal decision-making procedures for influencing policy. It has a constitutive aim, which is something like: to coordinate the beliefs and behavior of (generally speaking) ideologically aligned individual agents in order to promote a particular political agenda. But on a more general or broad picture of the Republican Party, for example if it were understood as a kind of collection of all the individuals in the United States who hold certain political beliefs and vote a particular way, then it would not count as a political agent since such a collection would not have a clear set of aims or a decision-making procedure.

[14] This definition of what constitutes a political agent is meant to include the production of policy and legislation as a sufficient condition, but not a necessary one. Entities like the "Supreme Court" do not directly produce legislation (although of course they are indirectly involved in evaluating the legality of certain legislation and either overturning or protecting it), but they still count as political agents.

agent's having "a will" to stand up and walk around has to do with (perhaps controversially) a motivational attitude of theirs, such as a desire or intention. But the idea of "the will" of that agent is more broad—it has to do with the agent's general ability or capacity to translate motivational attitudes (a desire or intention to stand up, beliefs about reasons to stand up, etc.) into action (standing up).

The idea of "political will," at least as the term is used in the media, is slightly ambiguous between these two concepts. Political will could reasonably be understood, for example, as something like a desire or wish for action on an issue in the public sphere, often in the form of policy or legislation. This is the view of political will that perhaps matches most clearly the way it is used by many politicians and news outlets when they claim there is a lack of "political will" to address climate change, for example. The problem is not that the requisite material resources and political avenues for action on the issue do not exist—rather, the problem is one of a lack of *desire* for action.

But equating political will toward some issue with a mere desire regarding that issue cannot be quite right. At least in the case of individuals, it is clear that a will cannot be understood as a desire alone, since individuals can desire all sorts of things that they make no effort to bring about. A will toward some action implies not just desire, but also a decision or judgment in favor of that action. Presumably, the same is true of collective agents. Further, in the context of collective agents, and political ones in particular, it is not obvious what would constitute a collective desire. For example, one option is that the desire is identical to popularity among "the people" ("the electorate" within democratic entities).[15] The people's aggregated desires could perhaps be informally understood as their will if, were they to be aggregated, the collective desires of the people form a simple majority in favor of some policy action. But it might also be formed through a more formalized process if the desires are in fact aggregated, for example through a vote or other procedure.

However, the mere fact of majority support for policy action on some issue (whether informal or not) does not appear to be sufficient for there to be political will that results in action on that issue. According to a 2019 Pew Research Center report, a majority of Americans believe that the United States is doing too little to reduce the effects of climate change.[16] But that support alone has not (at least at the time of the survey) been sufficient to realize action in the form of policies or legislation aimed at significantly reducing the effects of climate change. Thus, the idea that *a* political will toward addressing an issue is

[15] This view is similar to Rousseau's conception of the "general will," understood as the collective will of all citizens, although Rousseau's theory of the general will is obviously much more complex.

[16] Funk and Hefferon, "U.S. Public Views on Climate and Energy."

somehow reducible to the aggregate desires of the people alone (or the electorate, citizens, residents, etc.) regarding that issue does not seem plausible. A better candidate might be something like a joint intention. But understanding the political will to address some issue as being constituted by a joint intention among "the people" will not work, because the kinds of groups that are capable of action are those who meet a certain set of criteria in order to qualify as a group agent, and "the people" do not fit this criteria.

One upshot of the preceding discussion is that the difficulty in determining what constitutes a particular political will to act is due in part to the fact that a political agent like the United States is a complex amalgamation of policymakers and institutions. To say that the United States lacks the political will to address climate change, even as a majority of individual constituent members (including policymaking members) may individually desire action to address climate change (or even specific policy proposals to do so), suggests that a broader understanding of the governing collective involved is necessary, one that includes the subgroups of relevant policymakers along with institutional players like the Supreme Court, and other institutions with decision-making procedures.

With this broader understanding of the United States in mind, a return to the other way of understanding "the will" seems appropriate. In other words, it leads us back to a consideration of the way an entity like the United States government processes the relevant inputs (constituent and policymaker desires and intentions, among other things) to produce an output in the form of action. In what follows I will therefore mostly talk about the will of political agents in a kind of old-fashioned way that may come apart from the concept of "willing" or "intending" as typically discussed in both the literature on individuals and on joint intention.

We can therefore understand "the will" of political agents as a power that political collectives have to move themselves to act, where that action often (but not always) comes in the form of things like the creation or evaluation of policy and legislation. Political will governs the process of trying to bring about a change to the political status quo. In most instances, this process involves deliberation among at least some members of the political collective, and results in something like an intention to act. This intention is not itself identical to the will; rather, the system itself is the will, and the intention to act (which may or may not actually result in action) is a product of this system. Further, any intention that results is not always reducible to the intentions of particular constituent group members, or even the aggregate intentions of all the group members. Political will is a kind of *inertia* toward change, which is governed by the system through which the collective wills itself into new forms.

Because political will can therefore be understood more generally as the process by which governments and institutions decide on and attempt to bring about action, a characterization of the will of a particular political agent depends on the governing and institutional structure of that political agent. In other words, the process by which a direct democracy moves itself to act will be distinct from the process by which a military dictatorship may move itself to act. There may be commonalities that can be generalized into an account of the will of political agents broadly construed, but for now I will focus on the will of democratic agents, using the United States as a specific example.[17]

3. WEAKNESS OF WILL AND PRACTICAL REASON

On the most common characterization of *akrasia*, or weakness of will, as it appears in individuals, an agent's will is weak if that agent intentionally acts counter to their own better judgment.[18] In other words, the agent believes there to be another course of action available to them that would best accord with their considered judgment; they simply do not take that course of action. The weak-willed agent acts intentionally in the sense that they have *some* reason for doing what they ultimately choose to do. They just do not perform the action that they take themselves to have more or most reason to do.

With respect to *group* weakness of will specifically, some theorists have argued that group *akrasia* can arise because of conflicts between the rationality of individual group members acting in light of their own interests as individuals and the rationality of the group decision-making apparatus. For example, using a version of the discursive dilemma, Pettit shows that groups can be either responsive to the views of individual members, or collectively rational (depending on their organization and deliberative process), but not both.[19] Pettit and List argue that it is because of this discrepancy between individual judgments and group judgments that group *akrasia* is possible—because "individuals face conflicts between acting for the group and acting for themselves."[20]

Conflicts between individual beliefs and group rationality can thus lead to one kind of akratic break. We can redescribe this conflict along the same lines

[17] I focus on the United States as an example because it is the system of government and history of climate policy that I am most familiar with. However, it is clear that focusing solely on one government is a limitation of this paper. Future work would benefit from expanding the analysis to other countries beyond the United States as well as to other political agents that are not countries.

[18] See, Davidson, "How Is Weakness of the Will Possible?" 21–42.

[19] Pettit, "Akrasia, Collective and Individual," 68–96.

[20] List and Pettit, *Group Agency*, 200.

used to describe the discursive dilemma's conflicting conclusion-oriented and premise-oriented forms of deliberation. If the individuals are acting for themselves or making judgments as individuals, then they will be making an inference ("no pay sacrifice") on the basis of one set of premises. The group will collectively make an inference ("yes pay sacrifice") on the basis of a different set of premises. Both the individuals *qua* individual and the collective *qua* group agent are operating on reasons and can be acting rationally. If the decision about what to do is reached based on the inference derived from individual members' judgments, then the group may not act in accordance with the evaluative conclusions drawn as a collective. This is one plausible way that group *akrasia* may arise.

But Pettit and List's view is too narrow to account for the variety of ways group *akrasia* can arise. Assuming that the only way that group *akrasia* might arise is through a divide between group rationality and the views of constituent members locates the akratic break not in the group agent itself, but only in the *connection* between group and individual members. But thinking back to the reasons Pettit's discursive dilemma gives us for talking about collective group deliberation more generally, why not think the group *itself* can suffer from internal akratic breaks? In other words, group *akrasia* does not just occur when there is conflict between individual rationality and group rationality, but might also occur from an *internal* failure of collective rationality or breakdown of will. In the remaining sections of this paper, I aim to provide an account of exactly how this might occur.

4. A MORE ROBUST ACCOUNT OF WEAKNESS OF POLITICAL WILL

When an agent engages in practical reasoning or deliberation, that agent is engaged in reasoning about *what they ought to do*. But within the process of reasoning about what one ought to do, distinctions can be drawn between the steps involved: for example, between the formation of intentional states on the one hand and the application of these intentional states on the other.[21] This can help us see why Pettit and List's account is too narrow to explain all cases of group *akrasia* (and therefore political *akrasia*). Pettit and List argue that group *akrasia* arises because of a tension between acting on the interests of individual group members or the collective (as derived through the collective decision-making procedure). In this kind of case, individual members would be forming different intentional states than the collective as a whole would.

But there may be other points within a collective agent's deliberation about what it ought to do beyond the formation of intentional states where an agent

[21] Heath, "Practical Irrationality and the Structure of Decision Theory," 251–73.

might be irrational, such as in the production of beliefs or attitudes, production of desires, or practical reasoning from motivational attitudes to a conclusion about what one ought to do. Theories that do not distinguish between the rationality of the content of an agent's motivational attitudes, the rationality of the agent's judgment about what would be best in light of those motivational attitudes, and their subsequent decision about what to do in light of that judgment cannot adequately capture all of the ways in which an agent may experience a breakdown of will.

The more expansive account of weakness of will from Amelie Rorty, on the other hand, offers insights into the way these distinctions come to bear in the case of collective agents.[22] According to Rorty's view, we can distinguish among at least five different factors relevant to an agent acting. These are:

1. An agent's general beliefs about appropriate human aims
2. An agent's commitment to actualize or realize those aims (in other words, to instantiate them in action)
3. An agent's interpretation of a particular situation
4. An agent's forming the intention to act
5. An agent's acting according to their decision

An agent can therefore suffer from an "akratic break" when the connection between any one or more of these factors breaks down in some way.[23]

In much the same way, a political agent's power or capacity to move itself to act through a system of practical deliberation is subject to many constraints that may cause a breakdown of will somewhere along the process. Such breakdowns might occur at various points in the process of trying to act, including in the formation of something like a collective desire or intention among the people or particular subgroups of people (such as the Senate); in deliberation by relevant subgroups or among the collective as a whole regarding competing proposals and courses of action; in the actual attempt to act by (for example) passing legislation or enforcing a policy. In what follows, I will categorize these

[22] Rorty, "Where Does the Akratic Break Take Place?" 333–46.

[23] For example, the first kind of break, which she calls *akrasia of direction or aim*, occurs between an agent's general beliefs about what is good (their "general principles and ends") and their commitment to using those evaluations to guide their actions. In such cases, the agent simply fails or refuses to commit to following what they judge to be best. This failure or refusal is itself a breaking down of the will. Similarly, an agent might experience *akrasia of interpretation* (a break between their general principles or aims and their interpretation of the particular situation they are in), *akrasia of irrationality* (a break between their personal evaluation or interpretation of a situation and their forming an intention or decision to act), or *akrasia of character* (a break between their decision and their behavioral actions). Rorty, "Where Does the Akratic Break Take Place?" 335.

failures according to the location of the break within the process of practical deliberation, using Rorty's analysis of individual *akrasia* as a model for understanding these breaks in the case of political agents.

But because the process by which political agents come to act is different in some relevant ways from the way an individual might come to act, any account of the kinds of akratic breaks experienced by political agents must attend to these distinctions. For example, it is not obvious that political collectives can have beliefs in the same way that individuals can; it is therefore not clear whether political collectives can experience weakness of will due to a break between their general beliefs about appropriate collective aims and their commitment to instantiating those aims by acting.[24] For this and other reasons, we can modify Rorty's original analysis and recharacterize the factors relevant to a political agent's acting as follows:

1*. The political agent's constitutive aims
2*. The political agent's commitment to realizing those aims
3*. The political agent's option set
4*. The political agent's judgment or formation of a formal intention to act
5*. The political agent's acting according to their judgment or decision

We can subsequently identify at least four points at which a collective political agent can experience breakdowns of will. These are:

a. Failure of commitment: a break between the political agent's constitutive aims and its commitment to realizing those aims
b. Failure of rationality: a break between the political agent's commitment to realizing its constitutive aims, and its interpretation of the circumstances or option set
c. True *akrasia*: a break between the political agent's judgment about what it ought to do and its decision to act on that judgment
d. Failure to follow through: a break between a political agent's decision about what to do and its actually bringing about its intended action

Though slightly modified, these are very similar to Rorty's original characterization of the breaks as they occur in individual agents. This should not be surprising given the analogy between individual agents and collective ones; their process of deliberation will be distinct, but should follow roughly the same

[24] There are many theorists who argue that groups can have beliefs or belief-like states, however: see for example, Gilbert, "Modelling Collective Belief," 185–204; Tuomela, "Group Beliefs," 285–318; and Lackey, "What Is Justified Group Belief?" 185–208. Though I take the opposite position for simplicity's sake, the account of collective political *akrasia* could easily be modified to accommodate a more robust view of group belief.

format. In the following subsections, I will further explain these breakdowns of will, using the United States' weakness of will with respect to climate change as an example throughout.[25]

4.1. Failures of Commitment

Collective political agents have particular constitutive aims, some of which may be formally demarcated in a literal constitution, but many of which may not be. A collective agent may fail to commit itself to realizing those aims.[26] One reason this might occur is that the agent fails to properly grasp its own constitutive aims; in other words, political agents might be mistaken about the constitutive aims they take themselves to have (or not have). But another, perhaps more common reason that a failure of commitment might occur is that the political agent is radically conflicted between two courses of action, one that accords with their constitutive aims and one that may not.

To see how, first consider how this might arise in an individual. An individual agent may be unable to commit themselves to act in accordance with their general aims as a result of radically conflicting desires or beliefs—they may be frozen or indecisive in the face of proposals for action that are mutually exclusive. In such cases, the agent takes the reasons for either proposal to be sufficiently motivating, yet they are still indecisive and ultimately unable to act given the conflicting nature of the proposals.

[25] While my primary example is the United States' inaction on climate change, it should be noted that this theory of weakness of political will is not meant to apply solely to inaction on climate change. Climate change simply happens to be a particularly interesting and compelling example, especially given the frequent tendency of the media and politicians to blame inaction on "a lack of political will."

[26] Though the group primarily under consideration in these examples (the United States government) will have constitutive aims that are identifiable in virtue of the literal United States constitution (for example, "establish Justice," "insure domestic Tranquility," etc., are directly stated in the preamble of the Constitution), this is not the case for every political entity. Further, we can derive the aims of the specific subgroups of the United States government by appealing to these larger aims. Given its shared role in helping bring about things like "establish Justice," we can say that the United States Senate has similar or subsidiary aims, which may be tailored in specific ways given the rules and powers that the group itself has. In cases where political groups do not have a literal constitution with aims specifically spelled out, we can still derive constitutive aims by thinking about the reason or purpose for which that collective has been formed. A base kind of constitutive aim for many kinds of democratic political groups, for example, might be to coordinate behavior and settle disagreements. Ultimately my goal here is not to argue for one set of constitutive aims for each kind of political group but to show how these might plausibly be understood.

Similarly, group agents can experience this kind of radical internal conflict through political polarization as the attitudes of individual members or subgroups move to become more extreme. When members or subgroups move in opposing or different directions, this can serve as a constraint on the overall group agent's ability to make decisions and bring about action. Polarization may even occur as an inevitable by-product of the deliberative processes of the group itself.[27] The group's representative attitudes and intentions may move toward a more extreme version of its initial tendencies as a result of confirmation bias and the creation of echo chambers in group reasoning; but individual members or subgroups may also polarize away from one another as a result of similar mechanisms, resulting in radical conflict within a larger political agent. As individual members or subgroups become increasingly polarized through the normal course of group practical deliberation, factions may shift (rationally) toward opposing poles. While one might assume that differing opinions would lead to compromise, increasingly radical internal conflict has the same effect on group agents as it does on a single individual agent—it constrains the group's ability to act at all. Thus, much like individuals can become frozen in the face of conflicting desires (resulting in a failure to act on either desire), so too can groups default to "no action" as a result of internal polarization.

This internal conflict may occur within several different kinds of subgroups of political agents. At the individual level, constituents and individual policymakers may be sufficiently polarized such that they are unable to form a commitment in the form of aggregate desire or intention for action that would accord best with the overall collective agent's constitutive aims. For example, assume that one constitutive aim of a deliberative democratic government is to provide for equal representation in decision-making procedures. Individual constituent group members may have radically conflicting views about whether this aim is best accomplished through legislation that expands and protects voting rights, or through legislation that makes voting more difficult in the name of preventing voter fraud. Even if individual constituent group members agree about the nature of the group's overall constitutive aim, they may come to view mutually exclusive proposals as the correct means for realizing that constitutive aim. The collective fails to commit itself to realizing its aims when the result of this radical internal conflict is a failure to commit on the basis of *either* proposal, leading to a breakdown of will.

At the party level, political agents with competing political subgroups that are at least somewhat on equal footing (most liberal democratic agents) are

[27] Cass Sunstein, for example, argues that group polarization is the "conventional consequence" of deliberation within groups ("The Law of Group Polarization").

particularly susceptible to polarization. The two-party structure of the United States, for example, is such that competition between parties is polarizing and self-reinforcing. The use of the filibuster and other veto points by a polarized, radically conflicted Congress has resulted in a reduction of the "legislative productivity of Congress as a whole."[28] Further, because translating majority support into action requires major legislation, which cannot always pass on the strength of one side alone, a joint desire among a majority of the internal branches of the United States (for example, the House and the president) may sometimes not be enough for the system to result in action.

The result of polarization within the United States government as a political entity is not, therefore, more extreme outcomes in terms of policy or legislation, but either a failure to act entirely or policy "drift" in the form of unguided policy change.[29] This kind of inability to act is not obviously a result of the larger group agent's being practically irrational. Internally, the system of deliberation is functioning as it is set up to do. Intense competition for institutional control incentivizes subgroup members to postpone decisions and leave major legislation for later in order to avoid controversy in the lead-up to the next election.[30] The process of deliberation and internal party functioning are themselves operating rationally; the problem is that this process of deliberation results in increasingly radical internal conflict, with subgroup members on both sides offering reasons for mutually exclusive proposals for action. The United States, as a collective, is effectively frozen in the face of this radical internal conflict, unable to bring itself to act on either set of reasons to bring about either proposal.

Nowhere is this more clear than with respect to major climate change legislation in the United States. Intense competition between Republican and Democratic subgroups disincentivizes the kind of cooperation required to produce major legislation while reinforcing increasingly polarized attitudes toward policy proposals. The Republican Party thus has reason to be against certain legislative proposals even as other United States subgroup members (the Democratic Party, executive branch, etc.) have reasons in favor of those same proposals. The resulting internal conflict, while itself produced through rational deliberation, ultimately causes the United States as a whole to fail to act rationally in the face of the impending climate crisis, as "a government that cannot respond to emerging challenges such as global climate change puts its

[28] Lee, "How Party Polarization Affects Governance," 261–82.

[29] Lee, "How Party Polarization Affects Governance," 262.

[30] Lee, "How Party Polarization Affects Governance," 276.

citizens and the broader world at increased risk."[31] A political agent that does not protect its constituent members is failing to act in its own best interests. When this occurs because of polarization and radical internal conflict, the agent fails because it is too conflicted to commit itself to bringing about that which is in its best interest or accords best with its constitutive aims.

4.2. Failures of Rationality

Political agents may experience a break between their commitment to enacting policy or legislation that would align with their constitutive aims (both formal and informal), and their interpretation of the situation they are in and the possible options they have. The break can be understood in one of two ways. First, a political agent might simply be mistaken about the circumstances in which it finds itself, and therefore be mistaken about the set of policy or legislative options available to it. Second, a political agent may correctly countenance the circumstances in which it finds itself, but fail to reason appropriately about which available courses of action would best accord with its constitutive aims.

The latter case is simply a mistake of instrumental rationality. In the former case, a political agent may come to regard as infeasible the various policy proposals that would best accord with its constitutive aims. One way this might occur is when political agents experience adaptive preferences. Adaptive preference change occurs when an agent adapts their preferences to the feasible option set. In other words, adaptive preferences "typically take the form of down-grading the inaccessible options."[32] Adaptive preferences are a purely causal, nonconscious "mechanism for dissonance reduction that operates on the preferences by which options are graded"; they serve to make us satisfied with our feasible option set.[33]

Consider again the case of climate change. On the basis of current analysis, the United States (among other nations) needs to cut nearly all of its current emissions in order to stave off the worst of climate change. A recent special report on climate change from the Intergovernmental Panel on Climate Change (IPCC) states that we have reached almost 1°C of warming above preindustrial levels. While the effects of global warming are already being seen across the globe, the report details that remaining below 1.5°C of total warming is critical; remaining below 1.5°C of warming rather than below 2°C could mean an enormous difference for preventing or mitigating the absolute worst effects of global warming. The special report stresses that while there are pathways

[31] Lee, "How Party Polarization Affects Governance," 274.

[32] Elster, *Sour Grapes*, 120.

[33] Elster, *Sour Grapes*, 124.

that would keep warming at or below 1.5°C, such pathways "require rapid and far-reaching transitions in energy, land,... infrastructure (including transport and buildings), and industrial systems." Further, these transitions are "unprecedented" in scale, and would imply "deep emissions reductions in all sectors, a wide portfolio of mitigation options and a significant upscaling of investment in those options."[34] Basically, most of the world needs to take nothing short of drastic action, and the United States is no exception.

While the physical resources and pathways exist for the United States to take such action, it suffers a breakdown of will attributable to a disconnect between its commitment to realizing its constitutive aims and its adapted preference toward climate policy that is politically feasible but will not bring about the kinds of drastic change that would actually accord with its best interests. For example, the most ambitious general climate plan yet (the Green New Deal) was voted down in the Senate in what was essentially a political stunt.[35] Even top Democrats had been openly critical of the plan. The chances of congress passing the sweeping emissions reductions, decarbonization rules, and electrification efforts required to truly enable the United States to do its part to keep the world below 1.5°C of warming are almost nonexistent. Instead, many recent policy proposals have focused on smaller, more achievable goals. Some proposals have even shifted the goalposts, aiming the United States at targets that would help keep warming below 2°C or 3°C, as opposed to just 1.5°C.

These shifts in policy and legislative goals show the way that political agents can experience adaptive preferences in a way that constrains what they see themselves as able to do. In dropping their pursuit of a comprehensive climate action plan aimed at helping keep the world below 1.5°C in favor of "more realistic" smaller policies that may get the United States on track to help keep emissions below 2°C, United States policy preferences have adapted to the political infeasibility of more sweeping and radical climate change mitigation. The United States would not have preferred 2°C of warming if the political option set included keeping warming below 1.5°C. But given that the political option set barely includes 2°C, its preferences have adapted.

The main upshot is that these adapted preferences are a kind of response by political agents to perceived political infeasibility. The agent suffers a failure in the reasoning between facts about their political circumstances, the way those facts come to bear on the policy and legislative options available to it, and the way in which these relate to its constitutive aims. In other words, there is a

[34] IPCC, "Global Warming of 1.5°C."

[35] Green, "Democrats to Move on from Green New Deal."

breakdown within the rational connection between an agents grasp of its aims and its assessment of how it can best fulfill those aims.

4.3. True Akrasia

Political collectives may also suffer a breakdown of will when there is a divide between their judgment about what would be best and the formation of an appropriate intention to act in virtue of that judgment. In other words, even if the governing collective reasons correctly from its motivational attitudes and interpretation of its option set to a judgment about what it ought to do, the collective may still fail to form a commitment to act on that judgment in certain cases. This is because an agent engaged in rational deliberation will ultimately come to two distinct conclusions: a judgment about what would be best, and a decision about what to do in light of that judgment. Thus, even when a political agent arrives at a judgment about which of its available policy options would be best, it remains an open question as to whether the agent will act in such a way that accords with that judgment.

In the case of larger democratic political agents, for example, many such entities have built in veto points that serve to constrain the majority, which means that even if the agent judges (as measured through majority agreement) that a particular policy or legislative proposal would be best, the political agent is subject to minority constraints that may prevent it from following through on that course of action, thus constituting a breakdown of political will. In writing about the failure to address inequality in America, for example, Alfred Stepan and Juan Linz note that the United States political system has many electorally generated and constitutionally embedded veto players, where a veto player is an individual or collective whose agreement is necessary for a policy decision.[36] For example, the Senate and the House both function as veto players for the United States. Because the consent or approval of a veto player is necessary for some policy to move forward, the existence of (more) veto players will make it more difficult to "alter the political status quo."[37] Many political agents have at least two formal veto players that effectively serve as structural limitations on the motivation of political agents, thereby constraining their ability to act.

Further, there are other "constitutionally embedded features" of democratic political agents that can constrain the will of the majority. In the United States, for example, every state in the union has an equal vote in the Senate, regardless of population, and the Senate has more power than the House in spite of

[36] Stepan and Linz, "Comparative Perspectives on Inequality and the Quality of Democracy in the United States," 841–56.

[37] Stepan and Linz, "Comparative Perspectives on Inequality and the Quality of Democracy in the United States," 844.

being "malapportioned." [38] So even if a majority attitude exists toward realizing a specific set of actions through legislation, these features of the political agent serve to undermine its own will by allowing for strong minority constraint on what alterations can be made to the status quo. Further, as Stepan and Linz point out, while "all of these majority-constraining features are constitutionally embedded and could, in theory, be changed by amendments supported by exceptional majorities of citizens," the United States' constitutional structure "enables minorities to block such amendments with comparative ease."[39]

While Stepan and Linz are discussing inequality-inducing (or equality limiting) features of the United States, it is easy to see how their arguments can be generalized to show how such features might constrain the nation's will to act in other areas as well. Even strong public support and a majority of the constituent subgroups of the government being in favor of acting on something like climate change will not guarantee action given the structure and nature of the United States as a political agent. Veto points and other constitutionally or procedurally embedded features of the system of deliberation therefore serve as one important type of constraint on the will of political agents. In cases where veto points are used by the minority to ultimately prevent legislative or policy action on some issue (thereby preserving the status quo), even when there is a clear majority of policymakers and constituent support for that legislation, the political agent is unable to move itself to act according to its best judgment.

4.4. Failures to Follow Through

Finally, political agents may also be weak willed because of their inability to follow through on even the formal commitments they make. In other words, the propensity of political agents to revise their policy decisions prematurely can also constrain their will. This coincides with Richard Holton's view that the "central cases of weakness of will are best characterized not as cases in which people act against their better judgment, but as cases in which they fail to act on their intentions."[40] Weakness of will, on this account, is something more like failing to be resolute enough; it arises when "agents are too ready to reconsider their intentions."[41] Of course, there might be cases where we reconsider our intentions because we realize they were ill judged or that new circumstances now make them inappropriate, and these are clearly not instances of weakness

[38] Stepan and Linz, "Comparative Perspectives on Inequality and the Quality of Democracy in the United States," 845.

[39] Stepan and Linz, "Comparative Perspectives on Inequality and the Quality of Democracy in the United States," 846.

[40] Holton, *Willing, Wanting, Waiting*, 70.

[41] Holton, *Willing, Wanting, Waiting*, 71.

of will. Rather, weakness of will is a kind of "unreasonable" revision of our intentions in response to the pressure of contrary inclinations.

Political agents frequently experience revision with respect to their intentions. In some cases, these revisions may occur through the agent's typical process of practical deliberation, but the result may still be unreasonable. For example, democratic political agents experience changes in their internal structure at fairly short intervals due to electorally generated shifts in power, which constrain the ability of such agents to follow through on their policy decisions without major revisions. In 2015, the Obama administration signed the United States on to the Paris Agreement through executive action. The Paris Agreement was set up to only take effect when at least fifty-five nations representing at least 55 percent of global emissions had formally joined—that finally happened in October 2016, and so the agreement went into force in November. Less than one year later, newly elected President Donald Trump announced that the United States would formally withdraw from the agreement as soon as it legally could (which would be four years from the signing date). In 2019, the United States submitted formal notice of intention to withdraw. And now, of course, the United States has reentered the agreement after another electorally generated shift in its internal makeup.

The problem here is that the agent *failed to be resolute* in its intentions, thus leading to a breakdown of will toward climate policy. Political entities are particularly susceptible to this kind of breakdown because many of them often undergo radical changes in composition every few years due to elections and shifts in public opinion. And the effect of these changes in composition with respect to climate policy, reduction of greenhouse gas emissions, and environmental regulations more generally is clear. According to analysis from a recent *New York Times* article, the United States under the Trump administration "officially reversed, revoked, or otherwise rolled back" over eighty environmental rules and regulations.[42] In particular, under the Trump administration, the Environmental Protection Agency "weakened Obama-era limits on planet-warming carbon dioxide emissions from power plants and from cars and trucks.... At the same time, the Interior Department worked to open up more land for oil and gas leasing."[43] The article warns that that these rollbacks will significantly increase emissions over the next decade, among other things.

[42] Popovich, Albeck-Ripka, and Pierre-Louis, "The Trump Administration Rolled Back More Than 100 Environmental Rules."

[43] Popovich, Albeck-Ripka, and Pierre-Louis, "The Trump Administration Rolled Back More Than 100 Environmental Rules."

The United States as a political entity fails in these cases because it goes back on its original intentions too easily.[44] The problem here is that the political agent *failed to be resolute* in its intentions, thus leading to a kind of motivational failure in bringing about action on climate change. The agent is unable to capitalize on existing inertia even when formal votes are cast in favor of realizing particular policy outcomes. Environmental regulations or emissions restrictions that are only in effect for a few years are not effective in bringing about their intended outcome; a political agent that is too quick to roll these back or revise them cannot produce tangible effects.

This, however, raises an interesting question about weakness of will that arises within political agents due to a failure to be resolute. In some cases, agents revise their intentions for good reasons—in other words, they might revise their intentions because their original intentions were misguided, harmful, or malformed. If political agents try to institute policies that are difficult to roll back, that may actually be a bad thing in the case of policies that are harmful. Further, in the case of electorally generated shifts in power, where these power shifts occur in ways that align with the political agent's general aims and processes, it might appear that a revision of intentions could in fact be reasonable. And an agent that revises its intentions for good reason is not suffering weakness of will, but simply changing its mind on the basis of some justification.

This all depends on what we regard as an *unreasonable* revision of intentions. Here, thinking back to the extent to which an agent's intentions or judgments are appropriately or rationally tied to their constitutive aims is helpful. If an agent revises their intentions in such a way that prevents them from taking the course of action that would best accord with their judgments about how to bring about their constitutive aims, then their revision is unreasonable. In the case of United States climate policy, failing to address climate change is not in the best interest of the United States. Thus, if the United States rolls back its climate policies and legislation for no other legitimate reason (for example if the United States realized its constitutive aims would be better realized through a different set of policies), then it is acting unreasonably and exhibiting weakness of will.

[44] One objection that has been raised to this point is that, in fact, there are just two different agents here—the transition between administrations marks a transition to a new agent. I find this fairly implausible, however, since the agent in question is still the United States, even if the administration's personnel makeup is different. It seems perfectly reasonable to think that agents can change even large pieces of their internal makeup and yet still be the same agent.

5. IMPLICATIONS AND OBJECTIONS

These examples highlight four distinct kinds of weakness of political will, which are categorized according to where they take place within a political collective's deliberative process. The examples are not themselves fully exhaustive with respect to the phenomena that collective agents may demonstrate when experiencing weakness of will. But all failures of will on the part of the collective should be analyzable according to the preceding account, which provides a taxonomy of these failures based on how (or really when) they arise within the process of deliberation.

The broadness of this theory of weakness of political will raises a potential worry, however. You might wonder why we should think that weakness of will is to blame in these cases of failure, as opposed to some other explanation.[45] After all, the process of making changes to the political status quo is extremely complicated—what value does the concept of weakness of political will add to the explanation of why attempting to make these changes sometimes goes awry? Social scientists have put forward a number of alternative explanations for failures of collective action, which primarily involve game-theoretic accounts like prisoner's dilemmas. Why not think one of these can better account for what is truly going on in cases of a "lack of political will"?

To begin with, it is not clear to me that the exact phenomenon that game-theoretic analyses are often directed at are all that similar to the one I am trying to explain. But even assuming that it is the same phenomenon, recent discussions about the role of game theory and prisoner's dilemma modeling in economic analyses have brought into question the idea that game-theoretic models actually offer explanations at all. Some philosophers have argued that modeling human cooperation by using a prisoner's dilemma cannot offer true explanations of field phenomena because such models are overly simplified and idealized.[46] In part, this is because modeling a collective-action problem as a prisoner's dilemma requires that we assume the agents involved are perfectly rational players with perfect information, who go on to make rational choices. We can see this, for example, in Stephen Gardiner's description of the collective approach to climate change as a tragedy of the commons. One premise of Gardiner's argument states that "when each agent has the power to decide whether or not she will restrict her pollution, each (rationally) prefers not to do so, whatever the others do."[47] This kind of assumption is a necessary feature of

[45] Thank you to an anonymous reviewer for raising this objection.

[46] Northcott and Alexandrova, "Prisoner's Dilemma Does Not Explain Much," 64–84.

[47] Gardiner, "A Perfect Moral Storm," 400.

most game-theoretic models in order to explain the failure to address climate change by the global community.

But as many theorists have pointed out, it also oversimplifies the picture of any particular country's preference set. According to Matthew Kopec, for example, the assumption "is only guaranteed to be true if states value economic output in a strictly positive way," but "nations seem to care about many things besides merely increasing their economic output."[48] Similarly, Peter Wood notes that the game-theoretic model assumes that countries have clear preferences, which are usually based on the aggregate welfare of the countries' citizens. But "in reality, different citizens have greatly different preferences, and the decision making is based on a political process," which complicates the overly idealized picture of a country's preferences in such a way that game-theoretic analyses cannot account for.[49] Finally, consider the following from Linn Hammergren:

> Recent experience in the United States with proposals on a national health plan, NAFTA, and tax reform are relevant in suggesting how information overload, intentionally distorted messages, uncertainty, emotional reactions, and conflicting secondary interests make it difficult for public and elites to discuss their way to an acceptable solution to what all perceive as a problem. This is not to say that policy making is irrational, but rather that a strictly rational model oversimplifies the situation of both individual and collective actors.[50]

Even in the case where many (perhaps even the majority of) individual constituent group members are in agreement that there is some problem that needs to be addressed, this is no guarantee that the group can easily arrive at a kind of simplified preference set regarding possible solutions.

Even on a less complex picture of collective action among individual agents, evidence from behavioral economics suggests that agents do not operate perfectly rationally nor do they always avoid cooperation even when it is seemingly in their interests to do so, which means analyzing these cases according to game theory will not necessarily help us predict (or explain) the choices agents make. It seems wrong to assume it would be more helpful in cases where the preferences and decision-making processes are even more complex, as in the case of group agents like countries or political institutions.

[48] Kopec, "Game Theory and the Self-Fulfilling Climate Tragedy," 10.

[49] Wood, "Climate Change and Game Theory," 154.

[50] Hammergren, "Political Will, Constituency Building, and Public Support in Rule of Law Programs," 17.

Thus, the standard game-theoretic accounts from social scientists cannot provide us with the resources necessary to understand the particular kind of phenomenon I am interested in. When we think about why we often fail to bring about changes to the status quo, even when it would be in our interest to do so, surely a part of the failure can be attributed to malformed preferences or purely economic interests. But as Hammergren notes, part of the explanation is also "found in the difficulties of translating a general desire or even a specific plan into a concrete series of actions, each of whose parts must also be 'willed' into effect."[51] This reference to the difficulties of "willing" a plan into action is telling. Even when there exists a general desire, or even a generally supported plan of action, there remains some missing piece that limits the motivational capacity of the collective. My contention, then, is that the philosophy of action and moral psychology literatures contain the resources necessary for us to analyze this missing piece, or to explain what is really going wrong when collective agents seem to suffer these kinds of motivational failures.

Two additional implications are worth discussing here. The first is that identifying the exact way in which a collective experiences a failure of will can help determine possible solutions or preventive strategies. Agents may take on various strategies to prevent themselves from suffering from weakness of will in the future. These can take the form of precommitments or binds, but also changes in internal structure when necessary. But an agent that suffers weakness of political will due to a failure of rationality will have different sorts of precommitment strategies than one that suffers weakness of will due to true *akrasia*. An agent that frequently suffers from weakness of will due to failures of commitment, for example, may need to change its incentive structure to avoid running into the kinds of weakness of will pointed out by Pettit and List, which occur due to tensions between individual members and the collective reasoning apparatus. On the other hand, an agent frequently failing to change the status quo because of true *akrasia* may need to amend its internal deliberative process to get rid of certain veto points (like the filibuster) in order to truly precommit itself to avoiding the weak-willed option.

This brings us to the second implication, which is that many of the failures of will experienced by collective agents are a result of internal features of those agents themselves. At least in the case of liberal democratic agents, many features that lead to their weakness of will are literally constitutionally embedded into their structural makeup. In effect, they are a feature of the way these systems are set up, as opposed to a bug. Further, the propensity of democratic

[51] Hammergren, "Political Will, Constituency Building, and Public Support in Rule of Law Programs," 8.

political agents to experience these as constraints on their motivational capacities is not obviously a negative. A group agent such as the United States has vast power, and the desires and attitudes of the whole and of subgroup members have changed drastically over time. In some cases, it may even be a good thing that it is hard for the United States to will itself to act, as it may prevent the United States from acting rashly or in ways that ultimately are not in its considered best interests.

On the other hand, these motivational constraints serve as a kind of drag on the inertia of change, often inhibiting the agent from willing itself into new and more just forms. The United States deliberates, sometimes even forming intentions or resolutions to act, and yet still cannot bring itself to alter a political status quo in which many people continue to suffer injustice, oppression, poverty, etc. Further, as Lee pointed out, the agent is unable to act in the face of impending crises, leaving constituent citizens vulnerable to the risks of things like climate change.[52]

Harvard University
cflowerman@fas.harvard.edu

REFERENCES

Davidson, Donald. "How Is Weakness of the Will Possible?" In *Essays on Actions and Events*, 21–42. Oxford: Oxford University Press, 2001.

Doan, Michael D. "Climate Change and Complacency." *Hypatia* 29, no. 3 (Summer 2014): 634–50.

Elster, Jon. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press, 2001.

Epstein, Brian. *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. New York: Oxford University Press, 2015.

Funk, Cary, and Meg Hefferon. "U.S. Public Views on Climate and Energy." Pew Research Center, November 25, 2019. https://www.pewresearch.org/science/2019/11/25/u-s-public-views-on-climate-and-energy/.

Gardiner, Stephen M. "A Perfect Moral Storm: Climate Change, Intergenerational Ethics, and the Problem of Moral Corruption." *Environmental Values*

[52] I am extremely grateful to Sarah Paul for her thoughtful comments and advice on multiple drafts of this paper. I would also like to thank Harry Brighouse, Mike Titelbaum, Megan Fritts, and Marcos Picchio for their feedback at various stages of the project. I am also thankful for the comments from two anonymous reviewers at the *Journal of Ethics and Social Philosophy*, as well as reviewers from several other journals.

15, no. 3 (August 2006): 397–413.

Gilbert, Margaret. “Modelling Collective Belief.” *Synthese* 73, no. 1 (October 1987): 185–204.

Green, Miranda. “Democrats to Move on from Green New Deal.” *The Hill*, March 28, 2019. https://thehill.com/policy/energy-environment/436171-democrats-to-move-on-from-green-new-deal/.

Hammergren, Linn. “Political Will, Constituency Building, and Public Support in Rule of Law Programs.” Washington, DC: U.S. Agency for International Development, 1998. https://pdf.usaid.gov/pdf_docs/PNACD023.pdf.

Heath, Joseph. “Practical Irrationality and the Structure of Decision Theory.” In *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet, 251–73. Oxford: Oxford University Press, 2003.

Held, Virginia. “Can a Random Collection of Individuals Be Morally Responsible?” *Journal of Philosophy* 67, no. 14 (July 1970): 471–81.

Holton, Richard. *Willing, Wanting, Waiting*. Oxford: Oxford University Press, 2011.

IPCC. *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-Industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*. Cambridge: Cambridge University Press, 2022.

Isaacs, Tracy Lynn. *Moral Responsibility in Collective Contexts*. Oxford: Oxford University Press, 2011.

Kopec, Matthew. “Game Theory and the Self-Fulfilling Climate Tragedy.” *Environmental Values* 26, no. 2 (2017): 203–21.

Lackey, Jennifer. “What Is Justified Group Belief?” *Philosophical Review* 125, no. 3 (July 2016): 341–96.

Lee, Frances E. “How Party Polarization Affects Governance.” *Annual Review of Political Science* 18, no. 1 (May 2015): 261–82.

List, Christian, and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford New York: Oxford University Press, 2011.

May, Larry. “Collective Inaction and Shared Responsibility.” *Noûs* 24, no. 2 (April 1990): 266–78.

Northcott, Robert, and Anna Alexandrova. “Prisoner’s Dilemma Doesn’t Explain Much.” In *The Prisoner’s Dilemma*, edited by Martin Peterson, 64–84. Cambridge: Cambridge University Press, 2015.

Pettit, Philip. “Akrasia, Collective and Individual.” In Stroud and Tappolet, *Weakness of Will and Practical Irrationality*, 68–96.

———. “Groups with Minds of Their Own.” In *Social Epistemology: Essential Readings*, edited by Alvin I. Goldman and Dennis Whitcomb, 242–70. Oxford: Oxford University Press, 2003.

———. "Responsibility Incorporated." *Ethics* 117, no. 2 (January 2007): 171–201.

Popovich, Nadja, Livia Albeck-Ripka, and Kendra Pierre-Louis. "The Trump Administration Rolled Back More Than 100 Environmental Rules. Here's the Full List." *New York Times*, January 20, 2021. https://www.nytimes.com/interactive/2020/climate/trump-environment-rollbacks-list.html.

Rorty, Amelie Oksenberg. "Where Does the Akratic Break Take Place?" *Australasian Journal of Philosophy* 58, no. 4 (December 1980): 333–46.

Shapiro, Scott J. "Massively Shared Agency." In *Rational and Social Agency*, edited by Manuel Vargas and Gideon Yaffe, 257–93. Oxford: Oxford University Press, 2014.

Stepan, Alfred, and Juan J. Linz. "Comparative Perspectives on Inequality and the Quality of Democracy in the United States." *Perspectives on Politics* 9, no. 4 (2011): 841–56.

Sunstein, Cass R. "The Law of Group Polarization." John M. Olin Law and Economics Working Paper, no. 91, 1999.

Tuomela, Raimo. "Group Beliefs." *Synthese* 91, no. 3 (June 1992): 285–318.

Wood, Peter John. "Climate Change and Game Theory: A Mathematical Survey." CCEP Working Paper 2.10, 2011.

Journal of Ethics and Social Philosophy
Vol. 27, No. 1 · March 2024

https://doi.org/10.26556/jesp.v27i1.2926
© 2024 Author

PATERNALISM, SUPPORTED DECISION-MAKING, AND EXPRESSIVE RESPECT

Linda Barclay

Among those who work in public policy to advance justice for people with cognitive disabilities, it is widely argued that supported decision-making must replace surrogate, or substituted, decision-making. From a legal perspective, surrogate decision-making is often decried as a human rights violation. From a moral perspective it is said to be an indefensible form of paternalism. Supported decision-making is the alternative that avoids these legal and moral failings.

In this paper, I will focus primarily on the anti-paternalistic argument in favor of supported decision-making. I will begin in section 1 by discussing recent debates within the paternalism literature to clarify the distinction between surrogate and supported decision-making, a distinction that is often underspecified or unclear in the legal, advocacy, and policy literature. I will rely on a distinction developed by Daniel Groll to argue that supported decision-making should be understood as treating the will of the agent as structurally decisive, whereas surrogate decision-making treats it, at best, as merely substantively decisive.[1] With the distinction between surrogate and supportive decision-making clarified, I will then turn directly to my main argument. At the heart of the rejection of surrogate decision-making is the belief that such paternalistic action *expresses* something fundamentally disrespectful about those upon whom it is imposed: that they are inferior, deficient, or childlike in some way. Contrary to this widespread belief, I will argue that surrogate decision-making often expresses more respect for people with lifelong, "severe" or "profound" cognitive disabilities than does the adoption of supported decision-making.[2] Specifically, in section 2 I argue that in some cases supported decision-making can arguably express that people with cognitive disabilities

[1] Groll, "Paternalism, Respect, and the Will."

[2] The terms "severe cognitive disability" and "profound cognitive disability" are controversial and used very differently in different jurisdictions. There is no universal medical consensus on how they should be defined. There is certainly controversy in individual cases as to which category an individual may fit into. Rather than attempting to define

lack equal moral value. In section 3, I argue that supported decision-making for people with profound intellectual disabilities can arguably express that they lack complex and rich inner lives. In short, if our aim is to ensure our behavior and practices express respect for people with lifelong cognitive disabilities, then sometimes surrogate rather than supportive decision-making will be a much better option.

As this summary of my argument makes clear, I am concerned with the expressive dimensions of surrogate versus supportive decision-making. The expressive meaning of our actions (or omissions) matters, morally speaking. That view has had a profound influence on recent discussions of both paternalism and egalitarian political philosophy. For example, the recent focus on relational egalitarianism in political philosophy arose partly in response to the troubling expressive dimensions of more dominant distributive approaches to equality, particularly luck egalitarianism.[3] One of the examples that Elizabeth Anderson famously used to illustrate these expressive concerns was the distribution of health care resources. She claimed that many luck egalitarians would distribute health care resources for paternalistic reasons. These reasons, she argued, fail to express respect for the so-called beneficiaries:

> In adopting mandatory social insurance schemes for the reasons they offer, luck egalitarians are effectively telling citizens that they are too stupid to run their lives, so Big Brother will have to tell them what to do. It is hard to see how citizens could be expected to accept such reasoning and still retain their self-respect.[4]

Numerous philosophers working specifically on the issue of paternalism have recently developed arguments that resonate closely with these expressivist considerations. For example, Seana Shiffrin disputes that paternalism is about, or only about, an unjust interference with liberty. Rather, paternalism is characterized by the paternalistic motive: the distrust the paternalizer shows for the practical reasoning or will of the paternalized subject, and their belief about their own superior capacities in this regard.[5] Paternalism, many now say, is first and foremost a failure of respect, specifically associated with how the paternalizer regards the paternalized subject as inferior or deficient in some regard, or can be arguably taken to express such an attitude.

the terms, my specific examples will make clear the nature and extent of the disabilities I discuss.

[3] Anderson, "What Is the Point of Equality?"; Scheffler, "What Is Egalitarianism?"; and Wolff, "Fairness, Respect, and the Egalitarian Ethos."

[4] Anderson, "What Is the Point of Equality?" 301.

[5] Shiffrin, "Paternalism, Unconscionability Doctrine, and Accommodation."

The expressive meaning of our actions is also important in the sphere of disability, not least in the area of decision-making. This, of course, should surprise nobody: the disabled are a highly stigmatized social group and as such are routinely vulnerable to disrespectful behavior from others. Nevertheless, it will be argued in this paper that recent social and legal advocacy for people with cognitive disabilities has offered an oversimplified picture of how best to express respect for people with cognitive disabilities in the sphere of decision-making.

1. Surrogate versus Supported Decision-Making

People with cognitive disabilities have historically been subject to paternalistic guardianship and surrogate decision-making. Recent legal and political developments have put increasing pressure on the acceptability of surrogate decision-making and have instead demanded that it be replaced with supported decision-making in most cases.[6] Animating such demands are not only the great harms that have been inflicted on people through surrogate decision-making and guardianship arrangements, but also a rejection of the disrespect that is taken to be expressed by such arrangements—namely that some people have inherently deficient or inferior agential capacities.

Before being in a position to scrutinize such claims, it is necessary to get clearer about the exact difference between surrogate and supported decision-making. The distinction is not always as sharp as one might expect, for reasons I will explain in a moment. A discussion of some recent paternalism literature will enable me to clarify the core of the difference.

At its simplest, surrogate decision-making refers to a situation where a guardian is charged with making decisions for another person who is deemed to lack decision-making capacity. Within the policy-focused literature, different types of surrogate decision-making are usually distinguished from one another. Guardians can be legally charged with the responsibility to make decisions for another based either on best-interest standards or on the basis of what that person would have decided themselves (if they were not experiencing decision-making incapacity). I will say a little more about different types of surrogate decision-making in the next section, but the term refers to any situation whereby a guardian makes a decision for another who is deemed to lack decision-making capacity. In contrast, supported decision-making allows a person to make their own decisions. In recognition that some people may find making decisions more difficult, supported decision-making refers to various ways in

[6] Cf. Bigby, Whiteside, and Douglas, "Providing Support for Decision Making to Adults with Intellectual Disability"; Kohn, "Legislating Supported Decision-Making"; and Series, "Relationships, Autonomy, and Legal Capacity."

which the decision-maker can be supported: for example, others can help them understand complex information, or their options, or the risks and benefits attached to various options, and so on. Depending on the disability, supported decision-making can also help a person articulate their decision.[7] In short, the core distinction between surrogate and supported decision-making is that in one case a person does not get to make their own decisions and in the other case they do, albeit with some support.

While the distinction should be clear enough, in practice and advocacy it can get murky: often the issue of *who* makes decisions is conflated with other issues. For example, it is often assumed that supported decision-making respects the choices of individuals that surrogate decision-making fails to do, as evidenced by the central focus on "choice" in supported decision-making policy and academic literature.[8] However, the issue of respecting a person's choices is not the same as allowing her to be the decision-maker, as we shall see. Another common conflation occurs when it is claimed that supported versus surrogate decision-making is a difference between respecting choice as opposed to acting in a person's best interests.[9] But this obscures the fact that very often the best way to promote someone's best interests is to respect their choices and preferences.[10]

A distinction developed by Daniel Groll allows us to more carefully home in on exactly what it means to be the decision-maker required by supported decision-making.[11] Groll's distinction also allows us to disentangle what it means to respect the decision-making authority of a person from other issues with which it is often conflated, such as "respecting their choices." A detour into the recent literature on paternalism leads us to Groll's important distinction.

Paternalism has been most typically described as the interference in a person's liberty for their own good.[12] More recently, some philosophers have cast doubt on this account of paternalism.[13] One important reason is that there are common cases where a person is clearly subject to paternalism but does not

[7] Further details of how supported decision-making can work in practice for even the most profoundly disabled are discussed in section 3 below.

[8] Bigby, Whiteside, and Douglas, "Providing Support for Decision Making to Adults with Intellectual Disability"; and Office of the Public Advocate, "Supported Decision-Making in Victoria."

[9] Flynn and Arstein-Kerslake, "Legislating Personhood."

[10] Howard and Wendler, "Beyond Instrumental Value."

[11] Groll, "Paternalism, Respect, and the Will."

[12] Dworkin, "Paternalism."

[13] Begon, "Paternalism"; Groll, "Paternalism, Respect, and the Will"; and Shiffrin, "Paternalism, Unconscionability Doctrine, and Accommodation."

have their liberty interfered with. For example, Shiffrin argues that paternalism can occur through omission, as when *A* refuses to help *B* build a bookcase because *A* believes *B* too often asks for help and would be better off developing their own confidence and carpentry skills by building the shelves on their own.[14]

Such cases have contributed to an alternative, motive-based account of paternalism. Motive-based accounts identify paternalism not in the act of interference as such, but in the motive of the paternalist.[15] The paternalist, *A*, both distrusts the paternalized agent *B*'s judgment or will about her own good, and also believes that his own judgment is superior. Most of us share Shiffrin's view that there is a paternalistic motive at play in the bookcase example, even though *A* does not interfere with *B*'s liberty. Rather, what Shiffrin identifies as paternalistic is *A* substituting his judgment for *B*'s with respect to a sphere of decision-making that rightly belongs to *B*.

Exactly how to characterize the paternalistic motive is controversial. Shiffrin says *A*'s distrust of *B*'s judgment and a belief that his own is superior is paternalistic when it concerns matters legitimately within *B*'s control, whereas others count distrusting *B*'s judgment as specifically paternalistic when it concerns more narrowly a judgment about what is good for *B*.[16] These differences are not of particular relevance to the arguments of this paper. What is directly relevant is that all such accounts believe that the paternalist expresses something fundamentally disrespectful and perhaps insulting about *B*. *A* expresses disrespectful attitudes when he treats his own judgment about what is good for *B* as superior to *B*'s own judgment on the matter—attitudes, for example, that *B* is deficient or childlike. The central objection to paternalism on motive-based accounts is based on the value we place on treating others with respect and expressing respectful attitudes in our behavior. Paternalism, it is said, is first and foremost a failure of respect.

Complicating matters, Groll argues that the paternalist motive can be present even when a paternalizer acts in accordance with the will of the paternalized subject, because the paternalized subject wills it. His central example concerns Bob, who needs a percutaneous endoscopic gastrostomy (PEG), a type of feeding tube, but refuses to have one inserted. Bob's decision-making capacity is not in doubt. Now imagine the surgeon takes it upon herself to decide what she ought to do. She weighs the fact that Bob might die without the PEG, but she also weighs Bob's sincerely avowed desire not to have one. Taking into account all of these factors, she decides that it would be bad for Bob to have the

[14] Shiffrin, "Paternalism, Unconscionability Doctrine, and Accommodation," 2013.

[15] Begon, "Paternalism."

[16] Begon, "Paternalism."

PEG inserted as to do so against his wishes would cause far too much distress and anguish. So she tells Bob, "I have decided you won't have a PEG inserted because you don't want one." It certainly seems reasonable for Bob to be perturbed and indeed annoyed at this way of putting things. He is entitled to say to the surgeon, "What do you mean *you* have decided? It was not your decision to make!"[17] According to Groll, the surgeon has acted on a paternalistic motive even though she does what Bob wants, and does it because he wants it.

Why is the surgeon acting on a paternalistic motive? Groll suggests that Bob's will should be authoritative, or what he calls "structurally decisive." What this means is that we should understand Bob as issuing something with the force of an order when he says he does not want a PEG. Here is how Groll puts it:

> When Bob declares that he does not want surgery, his will is authoritative. This means that Bob, and in this case no one else, is the de jure ultimate decision maker in Bob's case. In other words, Bob's will grounds a legitimate demand that the surgery not be performed; he is effectively issuing an order that he not have the surgery, an order that only he is authorized to give. And it is in the conceptual nature of an order that it be treated as what I will call structurally decisive in determining what to do—it is meant to supplant the reason-giving force of other considerations not because it outweighs those other considerations but because it is *meant to silence, or exclude, those other considerations from the practical deliberations of the subject of the demand,* in this case Bob's doctor.... We can put this idea as follows: the force of the reason not to do the surgery that is grounded in Bob's demand is insensitive to considerations of Bob's good.... The normative force of Bob's demand is not properly assessed by determining what good (for Bob) comes from following it.[18]

Clearly the surgeon does not take Bob's wishes in the spirit of an order: rather, she weighs them highly, decisively so in the end, and decides to follow them because doing so is best for Bob. According to Groll, this is an instance of failing to respect Bob's will typical of the paternalistic motive we discussed earlier: the surgeon distrusts Bob's judgment about his own well-being and expresses this distrust when informing Bob that she has arrived at her own (superior) judgment about what is best for Bob, and decided accordingly.

Groll contrasts Bob's case with that of Carl. Carl also does not wish to have a PEG, but unlike Bob, Carl is deemed to lack medical decision-making capacity. The surgeon is aware that Carl's health will suffer seriously without a PEG

[17] Groll, "Paternalism, Respect, and the Will," 707.

[18] Groll, "Paternalism, Respect, and the Will," 700–701 (emphasis added).

but also that forcing an invasive medical procedure on Carl against his wishes can be extremely deleterious for his well-being. She also believes that though Carl lacks formal decision-making capacity, a good life for Carl includes the ability to direct his life according to his own wishes as far as possible. In the end, she decides that his overall well-being will be advanced by respecting his wishes. Even though Carl's wish not to have a PEG inserted is respected, his will is not treated as structurally decisive. While the surgeon takes into account Carl's wishes, they certainly do not silence other considerations concerning his well-being playing a part in her practical deliberations. Carl's will in this case is substantively decisive in the sense that his wishes end up carrying the most weight in the surgeon's deliberations. Groll argues that in this case there is nothing odd when she says to Carl, "I have decided not to give permission for the PEG because you don't want one."

The contrast between treating another's will as substantively versus structurally decisive allows us to home in more carefully on the distinction between surrogate and supported decision-making. It is clearly not helpful to contrast surrogate decision-making with respecting another's choices, because Carl's choices are respected even though he is subject to surrogate decision-making. *He* is not the decision-maker in this case, the surgeon is. Similarly, it is confusing to contrast respecting a person's choices with acting in their best interests, because the case of Carl demonstrates that we can respect a person's choices because it is in their best interests. Respecting a person's choices, in other words, can be done for paternalistic reasons. Carl's surgeon still exhibits a paternalistic motive. Her distrust of Carl's judgment and will drives the nature of her practical deliberations, and her belief that it is she, not Carl, who must be the ultimate decision-maker.

If the aim of supported decision-making is to avoid paternalism and the expression of disrespect it is said to entail, then supported decision-making is best understood as decision-making whereby the will of the subject is structurally decisive. Once a person has received sufficient support to make a decision, their decision should be treated as authoritative, as silencing others' practical deliberations about what they (those others) ought to decide. Those others have no decision to make. In contrast, surrogate decision-making refers to any situation where another person makes the decision for an individual: this can include surrogate decisions where the will of a person deemed to lack decision-making capacity is entirely discounted, but also those cases where it is treated as substantively decisive, as in Carl's case.

Within policy and legal literature, the archetype of "bad" surrogate decision-making occurs when the choices and preferences of the person subjected to guardianship are entirely ignored—they are not even consulted. In such cases, it is presumed that a surrogate decision-maker can reach a judgment about the subject's best interests without taking into account their preferences and values.[19] This is of course highly dubious in the majority of cases for a number of reasons, including those discussed by Groll: that very often what promotes a person's best interests is to respect their choices. Nevertheless, there is little doubt that historically this is precisely how surrogate decision-makers executed their role. To decide that a person lacked decision-making capacity was to assume their wishes lacked any kind of status or value. More recent "good" forms of surrogate decision-making—often referred to as substitute decision-making—explicitly direct surrogate decision-makers to make the decision the person would have made themselves had they not lacked capacity. For example, the UK Medical Capacity Act (2005) directs surrogate decision-makers to "encourage participation—do whatever's possible to permit or encourage the person to take part" and to "find out the person's views—including their past and present wishes and feelings, and any beliefs or values." The National Disability Insurance Scheme Act (2013) in Australia requires a guardian to "ascertain the wishes of the participant" in the insurance scheme even when they lack decision-making capacity. The recently updated Guardian and Administration Act (2019) in Victoria, Australia, directs surrogate decision-makers to "make a decision that gives all practicable and appropriate effect to the person's will and preferences, unless this would cause them serious harm." These examples are indicative of widespread changes from "bad" to "good" forms of surrogate decision-making that have occurred in countless jurisdictions.[20] For my purposes, however, they all count as forms of surrogate decision-making insofar as the person herself is not the decision-maker: her will is not treated as structurally decisive.

Groll assumes that it is appropriate that Carl is subject to surrogate decision-making given his cognitive disabilities.[21] However, as I have indicated, such a view is no longer widely shared among disability scholars and advocates who call for the (near) abolition of all forms of surrogate decision-making in favor of supported decision-making. To a large extent, these claims are bolstered by the social model of disability that claims that much if not all of the incapacity associated with cognitive disability is due to incommodious or

[19] Flynn and Arstein-Kerslake, "Legislating Personhood"; and Howard and Wendler, "Beyond Instrumental Value."

[20] Kohn, "Legislating Supported Decision-Making"; and Series, "Relationships, Autonomy, and Legal Capacity."

[21] Howard and Wendler, "Beyond Instrumental Value."

unjust social arrangements.[22] Within this framework, it is denied that there is a group of people who incorrigibly lack decision-making capacity. Rather it is believed that all (or very nearly all) people can achieve decision-making capacity and thus be accorded decision-making authority with the right kind of social accommodation—namely, decision-making support.[23] Those who advocate for supported decision-making for people like Carl do not imagine that we should assist Carl to work out and articulate his preferences in order that his guardian can make the decision that he would have made himself had he been able. The process of providing adequate support for Carl is to ensure that *he* can exercise his decision-making authority. So while "good" forms of surrogate decision-making may have been considered visionary only a decade or two ago, they too have been increasingly subject to criticism.[24]

Supported decision-making has been given a considerable boost by recent developments in human rights law. While Article 12 of the United Nations Convention on the Rights of Persons with Disabilities (CRPD) (2006) might seem to suggest that people can lose the right to make their own decisions in extremely limited circumstances, the United Nations Committee on the Rights of Persons with Disabilities (2014) maintained in a General Comment that in fact Article 12 prohibits surrogate decision-making in favor of supported

[22] Peterson et al., "Supported Decision Making with People at the Margins of Autonomy"; and Series, "Relationships, Autonomy, and Legal Capacity."

[23] Flynn and Arstein-Kerslake, "Legislating Personhood."

[24] Series, "Relationships, Autonomy, and Legal Capacity." An anonymous reviewer raised the question as to how these recent legal, policy, and advocacy claims are related to the traditional philosophical distinction between "hard" and "soft" paternalism. In contrast to hard paternalism, soft paternalism does not interfere with choices that are fully autonomous, but only those choices that are not. It might be assumed that such interference does not express disrespect. I agree with others who argue that disrespect for people's agential capacities is not confined to disrespect for their autonomy, and it is "agential capacities," not "autonomy," that I focus on in this paper: cf. Jaworska, "Respecting the Margins of Agency"; and Howard and Wendler, "Beyond Instrumental Value." The large body of literature on supported decision-making also focuses on respect for agential capacities, and for good reason: whatever the merits of the philosophical distinction between soft and hard paternalism, as a matter of law, policy, and practice, "autonomy" is not the standard used for paternalistic interference. Indeed, interfering with choices that fail to meet any such high bar would, in practice, be considered quite offensive by most of us (Begon, "Paternalism"; Wall, "Self-Ownership and Paternalism"). In practice, as opposed to "ideal philosophy," the focus has been on articulating much lower standards of "decision-making capacity." Many defenders of supported decision-making simply deny that there is some group of people who incorrigibly lack all decision-making capacity (except perhaps those who are permanently unconscious). I certainly agree that we can express serious disrespect for the agential capacities of someone who is not fully autonomous, but in this paper I argue that both surrogate and supported decision-making can express such disrespect.

decision-making. Most disability scholars agree that the CRPD calls for the abolition of surrogate decision-making, even as some legal experts express skepticism about states' likely willingness to do this.[25] A number of scholars have argued that surrogate decision-making is a violation of human rights and the "backbone" or "lynchpin" in the restriction or loss of various other rights.[26]

There are further reasons to be skeptical about the use and value of even "good" forms of surrogate decision-making, that is, surrogate decision-making that puts the will and preferences of the person concerned at the center. There can be no doubt that surrogate decision-making has been widely used where a person was in fact capable of making their own decisions, or would have been with appropriate support, including, not least, support for the development of agency at a young age. This has led not only to enormous frustration, but has robbed people of the opportunity to develop their agency, the exercise of which can boost self-esteem and the sense of personal well-being.[27] Surrogate decision-making has been regularly abused by surrogate decision-makers who misuse their power to promote their own interests as opposed to the interests, let alone the will, of the person they are supposed to be deciding for. Even when not intending to misuse their authority, surrogate decision-makers often fail to act in the interests of the person concerned, as their judgment is heavily clouded by their own interests, values, and beliefs.[28] Finally, the expressive dimension of denying an already highly stigmatized social group the right to make their own decisions about matters of personal and often intimate concern is thought to be morally troubling, to say the least. When such restrictions are enshrined in law and public policy the negative expressive force is arguably turbo charged.[29] It is these expressive concerns that are my focus in the rest of this paper. In contrast to prevailing opinion in disability scholarship and activism, I will argue in the next two sections that in some cases we can arguably express more disrespectful attitudes about people with severe, lifelong cognitive disabilities by adopting (or attempting to adopt) supported decision-making. I will argue that in some cases surrogate decision-making expresses more respect for people with severe or profound cognitive disabilities than does supported decision-making.

[25] Series, "Relationships, Autonomy, and Legal Capacity."

[26] Bach, "Inclusive Citizenship"; and Flynn and Arstein-Kerslake, "Legislating Personhood."

[27] Bigby, Whiteside, and Douglas, "Providing Support for Decision Making to Adults with Intellectual Disability"; and Series, "Relationships, Autonomy, and Legal Capacity."

[28] Bigby, Whiteside, and Douglas, "Providing Support for Decision Making to Adults with Intellectual Disability"; and Series, "Relationships, Autonomy, and Legal Capacity."

[29] Anderson, "What Is the Point of Equality?"

2. Expressively Disrespectful Supported Decision-Making and the Value of Disabled Lives

There are, of course, multiple ways in which we can express disrespect for others: expressing that others are deficient with respect to their agential capacities is but one domain. For example, Anne-Sofie Greisen Hojland argues that sometimes avoiding paternalism conveys objectionable attitudes such as neglect and indifference, among other things.[30] Focusing on state action, she argues that the state can fail to treat its citizens as equals not only by failing to express that their agency is equally valuable to that of other citizens, but also by failing to express that their interests count equally. She invites us to see that standing idly by when a motorcyclist is about to careen down a steep and narrow road in rainy weather without a helmet "arguably conveys an attitude of indifference to their strong interests."[31] If this is so, then she argues we would need some way to weigh the objectionable expressive content of paternalistic action against the objectionable expressive content of non-paternalistic restraint, for which she offers a number of criteria.

Similarly, Viki Møller Lyngby Pedersen argues that sometimes people who avoid paternalism problematically express insouciance or indifference to the plight of others, which can also be a clear expression of disrespect for their equal status.[32] She asks us to imagine Joe pranking Ben by telling him that he (Joe) will drink a cup of poison that will kill him. After discussion, Ben is satisfied that Joe is acting voluntarily so stands idly by while Joe drinks what Ben believes to be poison. Pedersen argues that Ben's failure to save Joe is a morally dubious expression of insouciance or indifference to the plight of Joe. After having drunk the harmless substance Joe says "Come on Ben! Would you really let me do that?"[33] We can clearly make sense of Joe's disappointment and his sense that Ben does not pay sufficient heed to the value of his life.

Neither Pedersen nor Hojlund argues that the morally troubling expressive meaning of avoiding paternalism always justifies acting paternalistically. They agree that paternalism can also involve expressing problematic attitudes about a person's agency. Their main point is that both paternalism and refraining from paternalism can express problematic attitudes about others but that the literature on paternalism has exhibited a rather single-minded focus on the problematic expressive meaning of paternalistic behavior. Their position is

[30] Hojlund, "What Should Egalitarian Policies Express?"

[31] Hojlund, "What Should Egalitarian Policies Express?" 526.

[32] Pedersen, "Respectful Paternalism."

[33] Pedersen, "Respectful Paternalism," 430.

that there are a number of complex criteria that ultimately determine whether a paternalistic or non-paternalistic stance on each occasion expresses greater overall respect for the subject.

The arguments made by both Pedersen and Hojlund have particular force in cases of cognitive disability. I will develop this argument by discussing just one case with which I am familiar. Rose (name changed) was a fifty-year-old woman who developed severe lymphedema in her legs, making it difficult for her to walk and causing her serious pain. Rose was reluctant to seek medical attention, being terrified of doctors, although after receiving much support she agreed to do so. She was eventually diagnosed with lymphoma, a type of blood cancer. Rose refused any further medical treatment, even after a process of extensive support for her decision was provided. Numerous people close to Rose had conversations with her about the nature of her illness, what treatment would involve, and the consequences of not receiving such treatment. She remained resolute that she did not want treatment. When I spoke to Rose about her health, she told me that her legs were sore because every time she left the house people would shoot her in the legs. When I asked her if she would like to visit a doctor who could help her with the pain she told me adamantly that she did not like doctors and she just wanted people to stop shooting her in the legs. It was eventually decided by those involved with Rose's care that Rose had received extensive support for her decision and that it must be respected.[34] Rose eventually died from untreated lymphoma.[35]

Did treating Rose's will as structurally decisive express sufficient respect for the value of her life or for her equal moral status? To clarify, I am not asking whether Rose's will should have been treated as substantively decisive, such that out of concern for her well-being her refusal of medical treatment should have been respected. I will return to that question shortly. At this point I am only focusing on the fact that surrogate decision-making was rejected in favor of treating Rose's will as structurally decisive. As Groll puts it, this silences or excludes considerations of Rose's good or well-being playing a role in the

[34] Australia does not have formal or legislated supported decision-making, although, as noted above, much relevant legislation requires surrogate decision-makers to take into account or adhere to the will and preferences of the person. Most decision-making of this kind takes place on a very informal basis, between family and care providers. It is relatively rare for decision-making to be escalated to a formal authority and usually only happens when there is disagreement between informal parties. Rose's caregivers in this case believed they were morally and legally responsible for respecting Rose's will once she had been provided with extensive support.

[35] Lest this be dismissed as a bad example of supported decision-making, Flynn and Arstein-Kerslake, proponents of supported decision-making, explicitly defend respecting the life-ending decision of a person in just such a case as this ("Legislating Personhood").

practical deliberation of others. I will argue that we have reasons to believe that morally troubling attitudes were expressed about the value of Rose's life when her will was treated as structurally decisive.

Rose quite clearly did not show a strong appreciation of some of the salient facts about her illness. She believed her pain was caused by being shot in the legs. I do not believe she understood what lymphoma is. She had only a limited capacity to grasp what treatment might involve, partly because of her overwhelming fear of doctors and hospitals. It was very questionable that Rose fully understood either that she would die without treatment or what it means to die. Having known Rose, I do not believe any level of support would have helped her resolve these deep misunderstandings. Let us contrast Rose to the case of Joe. The way Pedersen tells the story, Ben scrutinizes Joe's decision to drink the "poison." He asks Joe why he wants to do this, he checks and double checks that Joe understands that the poison will kill him and that Joe fully appreciates the finality of what this means. Once satisfied that Joe really does understand what he is doing and what the consequences will be, Ben refrains from intervening out of respect for Joe's agency. Despite this, we are invited to consider whether Ben expresses a morally troubling level of insouciance for Joe's moral worth when he refrains from swiping the cup from Joe's hand. If we feel the pull of this concern, then it is magnified in Rose's case where we have clear reasons to believe that she had an insufficient grasp of the facts that bore on her preferences. To conclude that respect requires that others exclude considerations of Rose's well-being from their practical deliberations seems to me to betray a paltry idea of what respect for Rose requires.

Agency is not the only thing that determines our worth or standing and how we should be valued by others. Agency is one dimension of persons that should be appropriately respected: so too should their lives, and their important interests. This point should be felt *forcefully* by those familiar with treatment of people with disabilities. It is not only the agency of people with disabilities that is undervalued or denied; so too are the full range of their interests and even their lives, especially so for those with cognitive disabilities. The evidence shows that people with cognitive disabilities are often stripped of the right to make decisions about highly personal and intimate matters that they would be perfectly capable of making with adequate support. Equally, the evidence also reveals high rates of medical neglect, failure of basic accommodations, failure to provide safe, high-quality housing, radical social exclusion, and so on.[36] Once we acknowledge that very basic interests of people with cognitive disabilities have been

[36] Baladerian, "Sexual Abuse of People with Developmental Disabilities"; Horner-Johnson and Drum, "Prevalence of Maltreatment of People with Intellectual Disabilities"; Murphy and Bantry-White, "Behind Closed Doors"; and Troller et al., "Cause of Death

dismissed or discounted, including the interest in life itself, a single-minded focus on respect for agency is an oddly blinkered view about what we need to do to express fulsome respect for such people. Knowing as we do the history and ongoing contemporary evidence of the abuse and neglect of people with cognitive disabilities, it plausibly expresses morally troubling attitudes to treat their wills as structurally decisive when they make life-threatening choices, at least when their level of understanding remains very poor, despite extensive support. It is at least not implausible to suppose that such an anti-paternalist stance in Rose's case expresses morally troubling attitudes about the value of her life.

I will discuss two objections to my argument that treating Rose's will as structurally decisive expresses morally troubling attitudes about the value of her life. The first objection rejects the expressive meaning I attribute to treating Rose's will as structurally decisive, and the second turns on the supposed negative consequences of *failing* to treat Rose's will as structurally decisive.

First, a critic might deny the particular expressive meaning I attribute to treating Rose's will as structurally decisive. The people who decided that Rose's will should be treated as structurally decisive might claim they did so solely out of a strong conviction that respect for agency is of paramount importance and that they neither held nor intended to express any other attitudes, least of all about the lower value of Rose's life. Respect for Rose's agency, they might maintain, required of them that they excluded facts about Rose's well-being from their practical deliberations.

This response raises questions about how we determine the expressive meaning of people's actions or omissions, for which I will offer two brief suggestions.

1. The expressive meaning of our actions is not determined solely by the attitudes we sincerely avow or intend to express.[37] People can act on attitudes that they are not even aware that they have as the wealth of discussion on cognitive processes like implicit bias and stereotype threat have shown. To deny this is to assert that a person's actions cannot reasonably be read as expressing his problematic attitudes about race just because he sincerely believes he holds no such attitudes, or that his actions cannot reasonably be read as expressing problematic attitudes about women just because he sincerely believes he has no sexist attitudes. To the contrary, people of good will who are open to their own fallibility are aware that it is sometimes our very actions (or omissions) that should alert us to the possibility that we *do* hold morally troubling attitudes

and Potentially Avoidable Deaths in Australian Adults with Intellectual Disability Using Retrospective Linked Data."

[37] Anderson and Pildes, "Expressive Theories of Law," 1513.

despite our sincerely held values. In light of the overwhelming evidence of how people with disabilities are treated in our society, it is more plausible than not to suggest that morally troubling attitudes about the value of the lives of people with severe cognitive disabilities are widespread. As such, it is not implausible to suggest that such attitudes are held and expressed by those who treat Rose's will as structurally decisive despite the extensive evidence of her limited levels of understanding. In any case, it is certainly not convincing to assert that no such attitudes are held or expressed just because the persons involved sincerely disavow that they hold such attitudes.

2. My second brief suggestion is to go further and deny that the meaning of a person's actions is solely determined by their attitudes (whether they are conscious of them or not). Here I follow Anderson and Richard Pildes, who assert that actions have public meanings.[38] Those who act in a certain way may not understand the public meaning of what they do, as when I hold up my middle finger to the face of another person believing that I am offering praise. This is a common enough occurrence when we are in an unfamiliar cultural environment. Indeed, Anderson and Pildes argue that the public meaning of an action is often not even determined by shared understandings of what it means. They offer the example of men complimenting women on their appearance in the workplace. Not long ago, few people recognized treating women as sexual or aesthetic adornments in the workplace as insulting.[39] But despite this meaning of the practice not being widely shared, that is indeed what it meant. The meaning of any action, according to Anderson and Pildes, is partly determined by how it "fits" with other practices and norms in the community: "Although these meanings do not actually have to be recognized by the community, they have to be recognizable by it, if people were to exercise enough interpretive self-scrutiny . . . a proposed interpretation must make sense in light of the community's other practices, its history and shared meanings."[40] What they convincingly suggest is that had the community engaged in interpretive self-scrutiny at the time they may have noticed that the practice contradicted norms of professional conduct among men, and also the various ways it slotted into the gendered hierarchy of labor, traditions of excluding women from positions of responsibility, and so on.[41]

In light of our community's other practices, our history and shared meanings around disability, treating the wills of people with severe cognitive

[38] Anderson and Pildes, "Expressive Theories of Law."

[39] Anderson and Pildes, "Expressive Theories of Law," 1525.

[40] Anderson and Pildes, "Expressive Theories of Law," 1525

[41] Anderson and Pildes, "Expressive Theories of Law," 1525.

disabilities as structurally decisive in life-threatening situations despite their clearly limited levels of understanding can be plausibly said to express troubling attitudes about the value of their lives and the weight we give to their pressing interests. Given the widespread disregard we have always shown toward the lives, well-being, safety, comfort, and security of people with cognitive disabilities, the meaning of treating their wills as structurally decisive in the face of life-threatening behavior can express troubling attitudes about the worth and value of their lives, even if some individuals who choose to treat such a person's will as structurally decisive hold no such attitudes.

A second objection to my argument that treating Rose's will as structurally decisive expresses morally troubling attitudes focuses on the purported negative consequences for Rose if we fail to treat her will as structurally decisive. Namely, it might be thought to follow from my argument that we should impose treatment on Rose against her will. That would be no small thing. Supposing that nothing we could do for Rose would resolve her terror of doctors and hospitals, that would likely subject her to distress. Moreover, the treatment for blood cancers is grueling by anyone's standards, involving months if not years of chemotherapy, radiation therapy, and sometimes surgeries. So apart from her terror, Rose would have to endure extensive physical, emotional, and social burdens. To inflict these on a person against her will seems unconscionable, even if we are right that her will is based on a serious misunderstanding of the basic facts. Am I really suggesting that forcing such treatment on Rose expresses greater respect for her than "respecting her wishes"?

I am suggesting no such thing. And talk of "respecting her wishes" is misleading. My concern about treating Rose's will as structurally decisive is based on the morally problematic meaning that is thereby expressed, not on the fact that she is allowed to die. It is based on the fact that treating Rose's will as structurally decisive is to ignore facts about her well-being for the purposes of our practical deliberation. As a *surrogate decision-maker* I might also "respect her wishes" and decide she should be allowed to die. It may well be that given Rose's intransigence, subjecting her to invasive chemical and radiation treatment would cause her intolerable levels of distress. Out of concern for her well-being I might decide to respect her wishes not to receive medical treatment. Considerations about her well-being, in other words, lie at the heart of my practical deliberations as a surrogate decision-maker. I would acknowledge that a tragic choice has to be made here: between taking action that fully respects the value of Rose's life on the one hand, and avoiding inflicting intolerable distress on her on the other. That is an honest appraisal of the nature of the tragic decisions that surrogate decision-makers must sometimes face. We express respect for Rose by acknowledging that we cannot act to preserve her valuable life without causing her unacceptable levels of distress. What does strike me as disrespectful is to deny that such tragic choices exist by conceiving of our duties to respect Rose as being exhausted by treating her will as structurally decisive, so long as we have provided her with extensive support for her decision, and irrespective of how much basic misunderstanding she continues to display.[42] Ignoring, or indeed refusing, to consider the well-being of a person in our practical deliberations when she remains deeply confused about matters of a life-threatening nature because of a cognitive disability is not a victory for expressive respect. As I have argued in this section, it is more plausible to suggest that it expresses morally problematic attitudes about the value of her life or her basic interests.

3. Expressively Disrespectful Supported Decision-Making and the Rich Inner Lives of People with Cognitive Disabilities

Rose was able to clearly articulate her preferences and more generally engage in fluent conversation with others. Some people with what are dubbed "profound" cognitive disabilities are not able to speak and apparently have very limited capacity to process language or to reason in ways we are familiar with. When people have lived their whole lives with such disabilities, we have little evidence that they are likely to have the complex beliefs and values that other people do, including people with less serious cognitive disabilities. How is supported decision-making supposed to work for them?

Supported decision-making, or something close to it, is possible for people with lifelong profound cognitive disabilities. Language is obviously not the only way that people can communicate with one another. All of us communicate extensively with gestures and sounds. Someone attentive to the communicative modes of a person who is nonverbal can often understand their wishes with respect to things like what they like to eat, whom they do and do not like living with, what activities they do and do not enjoy, which support workers they feel comfortable with and which they do not, and so on. With respect to most of these matters it should often be possible to treat the will of the person concerned as in some sense structurally decisive. If the person communicates that they do not enjoy a certain kind of food or certain music then in most situations the music should be changed and alternative food offered. Of course, such options will sometimes be more difficult when people live with others. The

[42] I suspect that many proponents of supported decision-making within disability activism believe that a person provided with high-quality support will not continue to hold false beliefs or deep misunderstandings. Of course, quality support for decision-making will help eliminate misunderstandings. But to suppose that support can always do this betrays a naive view about severe intellectual disability (or just human nature more generally!).

point is that, insofar as these sorts of complications are not relevant, the wishes of the person concerned can and often should be treated as structurally decisive.

However, the range of matters that impinge on the lives of all people, including those with profound cognitive disabilities, is extremely wide. It includes not only matters about what we like to eat, what music we like to listen to, whom we want to spend time with, and so on, but also includes what religious practices, if any, we might engage in, how finances should be managed, whether to undertake grueling medical treatment, and so on. Like anyone else, a person with lifelong profound cognitive disabilities confronts many complex matters that can involve quite dramatic risks and benefits, yet it may not be clear how we could come to know their will. A parent might deny her son a COVID vaccination on the grounds that, according to her, he does not believe in vaccination; a Jehovah's Witness might declare that her daughter does not want a blood transfusion; yet another denies that her son wants a PEG inserted because of his love of food, even though it leaves staff at his residential facility having to call an ambulance on a regular basis when he experiences life-threatening choking episodes.

The obvious question is: How can a person claim to know the will of the subject in these cases? We cannot even consider the possibility of treating the will of the person as structurally decisive without first having grounds to be confident that we know what their will is.

I admit that I am skeptical about our ability to know what a person's will is in many such cases, partly because I am skeptical about the capacity of a person with lifelong, profound cognitive disabilities to develop a will in complex cases of this kind. But these skeptical concerns are not my focus here. Rather, I will discuss a number of concerning moral consequences of trying to apply supported decision-making in such cases, including the unacceptable expressive dimensions that arise when we display too much confidence in our ability to know the will of people with profound disabilities. I develop these criticisms by discussing an account of supported decision-making articulated by Leslie Frances and Anita Silvers.[43] Other defenders of supported decision-making for people with lifelong, profound cognitive disabilities have gestured at how the process could work: for example, Eilionóir Flynn and Anna Arstein-Kerslake suggest that the facilitator's role is to "imagine" what the person's will and preferences might be.[44] Silvers and Frances are alone in offering a detailed and

[43] Frances and Silvers, "Liberalism and Individually Scripted Ideas of the Good"; and Silvers and Frances, "Thinking about the Good."

[44] Flynn and Arstein-Kerslake, "Legislating Personhood," 95.

rigorous account of how the process of support for decision-making should work for people with lifelong, profound cognitive disabilities.

Silvers and Frances develop their view in the context of criticizing liberal political theory. While endorsing liberalism's commitment to diverse individual conceptions of the good life, they criticize what they take to be a widespread assumption that each individual must develop her conception of the good independently of others. They argue that we should accept not only diversity with respect to conceptions of the good, but diversity in the process by which different individuals arrive at their conception of the good. As they put it, there should be tolerance not only about the substance of the good, but also about how it is formed. Liberalism, they claim, fails with respect to the latter because it demands that the "proper process for arriving at and articulating the good specifies that individuals make determinations of their good on their own."[45] This, they argue, necessarily leads to the exclusion of people who are heavily reliant on others for formulating and articulating a conception of the good.

How then do people with profound cognitive disabilities form a conception of the good? According to Silvers and Frances they can do so by deploying a prosthetic reasoner, whom they call a trustee. As they put it:

> We envision the trustee does not step into the subject's role in shaping a personalized notion of the good. Instead, as a prosthetic arm or leg executes some of the functions of a missing fleshly one without being confused with or supplanting the usual fleshly limb, so, we propose, a trustee's reasoning and communicating can execute *part or all* of a subject's own thinking processes without substituting the trustee's own idea as if it were the subject's own.[46]

Silvers and Frances are clear that they see trustees as facilitating a conception of the good for even the most profoundly disabled people, hence the explicit reference to a trustee possibly executing *all* of a person's reasoning processes.[47] They say that people who cannot use language and who we have good reason to suppose are incapable of most conceptualizations and reasoning can use a trustee as a prosthetic in this way. Appealing to ideas of relational autonomy, they argue that using a trustee to execute the subject's reasoning and communication is just a matter, to a more extensive degree, of the ways all of us rely on interactions with others to develop our conceptions of the good. Or as they

[45] Silvers and Frances, "Thinking about the Good," 477.

[46] Silvers and Frances, "Thinking about the Good," 485 (emphasis added).

[47] Perhaps they would exclude people thought to be "brain dead" or who show no demonstrable brain activity.

put it, "this prosthetic practice differs in extent and implementation, but not in nature, from commonplace social interactions that facilitate people's development of their notions of the good."[48]

Much of what Silvers and Frances say about prosthetic reasoning is at a very high level of abstraction, as these quotes suggest. What does it consist in, exactly? I take it that they are not suggesting that trusteeship involves merely being responsive to a person's unique way of communicating, and translating such communication into a form that others can also understand. If that is all they mean, then talk of a prosthesis seems entirely out of place. None of this common practice requires a prosthesis to *execute all of the reasoning and communicating* of the person—it does not require a prosthesis at all. It requires only the existence of others with a close relationship with the subject, who understand her way of communicating, and who have a deep commitment to ensuring her will is acted upon. This is a commonplace activity in high-quality relationships with people with profound cognitive disabilities.

Therefore, I assume Silvers and Frances have something more ambitious in mind: that the prosthesis's unique role will be to execute the reasoning of the subject in forming a broader conception of the good, one that reaches beyond that much more limited range of matters that the person has clear preferences with respect to and can communicate herself (to those who understand her). One possibility here is prosthetic reasoning as a kind of *extrapolation*: that the trustee reasons on behalf of a person that because he has a great love of food, he rejects a PEG, or because he dislikes needles, he rejects medical treatment. Yet it is clear that these conclusions cannot be straightforwardly extrapolated from a person's limited preferences about food and needles. I too love food and hate needles; nevertheless, when push comes to shove, I would almost certainly reevaluate or just dig deeper into aspects of my conception of the good to accommodate my changed circumstances. Similarly, Silvers and Frances make some rather oblique references to the connection between conceptions of the good and social scripts.[49] It is true that chunks of our conception of the good come from the social roles we inhabit: parent, teacher, Muslim, and so on. But very few of these social roles will be so tightly scripted so as to dictate clear answers to many of the quandaries that frequently arise in our lives, including the lives of people with profound cognitive disabilities, such as whether to accept a PEG. Moreover, the law and morality do not typically permit us to impose life-threatening or even life-changing aspects of social scripts onto people unless they have endorsed them, or at the very least not rejected them:

[48] Silvers and Frances, "Thinking about the Good," 495.

[49] Frances and Silvers, "Liberalism and Individually Scripted Ideas of the Good."

Jehovah's Witness parents do not have authority to deny their young child a blood transfusion, nor do members of religious groups have authority to marry off their young daughters. In cases such as these, we will have no evidence whatsoever as to whether a person with profound disabilities has endorsed, or merely rejected, such aspects that others claim are part of their socially scripted good.

What else, then, might prosthetic reasoning consist in, if not merely attending to what the person communicates about her likes and aversions, or straightforward extrapolation from such? It seems as though we are being invited to take a leap of faith: to accept that the reasoning conducted by the prosthesis on these complex issues is really the subject's own. It will involve sensitivity to the expressed wishes of the subject, and a degree of obvious extrapolation, but will clearly involve much more as well. We should accept that the "much more" really is the subject's own, when the reasoning is conducted by a diligent trustee. It is this ambitious idea that seems to make the most sense of the idea of a "prosthesis."[50]

There are skeptical questions to raise here, clearly. As others have commented, prosthetic limbs do not have minds of their own, a key difference that raises genuine concerns about how someone executing all of the reasoning for another can exclude her own reasoning from the process, or even distinguish between her reasoning and the subject's.[51]

I want to sidestep these skeptical questions in order to home in on moral, rather than epistemic, concerns. For the sake of argument let us take the leap of faith and accept that a diligent trustee *can* execute the functions of another person's mind as deeply as prosthetic reasoning seems to entail. Would it be morally acceptable to do so? Most of us would forcefully reject someone presuming to take on such a role with respect to our own minds. Indeed, we very actively limit others' access to our minds. A degree of opacity, concealing large swathes of our inner lives, seems to be a basic need. Many of our desires, values, preferences, hopes, fears, and passions remain private, or are revealed only to some, when we deem it appropriate or desirable to do so. Moreover, when we do reveal information about our preferences and values, we do so with a degree of authorial control: we tend to carefully curate the way we present information

[50] In a recent article, Leslie Frances uses multiple examples of prosthetic tools and "guardrails" that do not seem to have much relevance to people with profound cognitive disabilities, which is our focus here. She offers examples of text reminders, automatic bill payments, automatic delays for large expenditures, the use of financial advisors, and so on. The cases Frances discusses where these prostheses and guardrails can be deployed only reasonably concern people with less severe cognitive disabilities actively wanting to manage their own financial affairs, albeit with some support ("Supported Decision-Making").

[51] Series, "Relationships, Autonomy, and Legal Capacity."

about our inner lives, for example by presenting a particular narrative about the origin or reason for some of our desires and values. None of us wants to be fully laid bare and it is highly unlikely our sense of dignity and self-worth would survive such exposure.

Does the ambitious idea of prosthetic reasoning as envisaged by Silvers and Frances appropriately respect the importance of opacity? It does not seem to. Someone able to execute all of the reasoning for another subject must be presumed to have very deep access to the subject's mind: it is not even clear whether on their account there are distinct minds to talk of. In any case, I assume they hold that such extensive prosthetic reasoning must be deeply informed by knowledge of the subject's desires, preferences, values, fears, and pleasures. The subject seems to have lost opacity altogether on this account.

Apart from the subject's dignity, there are other reasons to value opacity that are connected to vulnerability, and some of these reasons apply just as much to people with profound cognitive disabilities as to other people. We are vulnerable to anyone who is confident that they have unfettered access to our minds such that they can execute its functions. In such circumstances, the threat of inappropriate behavior if not outright abuse looms. If we imagine a case where the subject later experiences an improvement in her cognitive abilities it would be incoherent for her to claim that the decision made earlier was not her own, or even one she did not endorse.[52]

These moral concerns about the idea of prosthetic reasoning are turbo charged by noting the subject's lack of control and authorization over "their" prosthesis. The runner exerts control over her prosthetic leg in a manner that is clearly disanalogous to the control a profoundly disabled person exercises over "their" prosthetic reasoner.[53] There is an obvious sense in which the runner authorizes the prosthetic limb to execute the function of running: she chooses to fix it on before she runs the race. How does the profoundly cognitively disabled person authorize or reject "their" prosthetic? What are the grounds on which we can be confident that her authorization has been provided?[54] It may

[52] Series, "Relationships, Autonomy, and Legal Capacity." Indeed, a number of legal commentators have recently argued that supported decision-making can and has been misused in ways that bear striking similarity to the more familiar abuses of surrogate decision-making, and that there are aspects of various legal regimes that lend support to this problem (see Kohn, "Legislating Supported Decision-Making")

[53] Wasserman and McMahan, "Cognitive Surrogacy, Assisted Participation, and Moral Status."

[54] It is worth pointing out here that virtually all proponents of supported decision-making explicitly require that the person relying on the support selects and authorizes a support person or persons: cf. Bach, "Inclusive Citizenship"; Flynn and Arstein-Kerslake, "Legislating Personhood"; and Series "Relationships, Autonomy and Legal Capacity." This

be that in some cases she can clearly express her rejection of a prosthetic reasoner and communicator by expressing distress or rejection when the prosthetic attempts to engage with her. But presumably in more cases than not the person may express very little in this respect, so the question remains: What is our evidence that the person has authorized this prosthesis to access and execute the functions of her mind? When a full prosthetic reasoner takes on this role, with little to no evidence that the person controls or authorizes the process, then the attitudes being expressed about that person are troubling: that her authorization and control is not required before he, the trustee, presumes to enter her mind and execute its functions. It bears repeating: those of us without profound cognitive disabilities would never accept another person adopting such a role with respect to our own minds.

While these concerns are important, I think there is something more directly troubling with the idea of prosthetic reasoning. Up until this point, I have assumed for the sake of argument that a person *can* access another's mind to the extent that prosthetic reasoning seems to presuppose. But this assumption itself raises serious moral questions. Imagine a person, Ken, who takes it upon himself to speak for his partner whenever he can. In a range of professional, health, and social settings, he confidently tells others what her preferences and values are and therefore what she would like to be done as it concerns her own good or well-being. I suggest our indignation at Ken's behavior is not exhausted by the fact that he violates her privacy with respect to her own mind and renders her vulnerable to any misuse of the role he has taken upon himself to play. In addition, I would suggest that we might take offense at Ken's very assumption that he has the level of access to her mind that he claims to. Specifically, Ken seems to express morally troubling attitudes that his partner lacks a deeply rich and complex inner life that, by its nature, would render his access to it extremely limited. We might say that Ken fails to respect his partner as a separate person. I mean this not in the sense that Rawls did—namely, as a criticism of utilitarianism for trading off some individuals for the overall good—but rather, the sense in which Ken fails to appreciate his partner as a separate person refers to his failure to appreciate that she has a rich and complex mental life that is barely accessible to him. What access he does have should always be tempered by a respectful acknowledgement of how incomplete it is, and how any beliefs he has about her will are likely to be partial and often just wrong. Without this recognition and acknowledgement on Ken's part, why need he bother to wait for his partner to speak for herself? There may well

of course just raises the question of how this is to be secured in the case of people with lifelong, profound cognitive disabilities. There is little consistency on this point in the legal and policy literature on supported decision-making.

be reasons to do so, but that she knows her own mind far better than he could ever hope to would not be one of them.

I believe that the idea of prosthetic reasoning for people with profound cognitive disabilities fails to express appreciation for the subject as a separate person, one with a rich and complex inner life that is not simply there for a diligent trustee to access so as to execute its various functions. We express respect for other persons when we acknowledge this, and thereby concede that we have at best very limited access to their inner lives. Such epistemic humility directly expresses our appreciation for the rich and complex inner lives of people with profound cognitive disabilities despite their cognitive limitations, and an acknowledgement that they share this feature with all others. As such, their inner lives are no more accessible to us than Ken's partner's is to him.

It is important to labor this point because this is exactly the kind of respect that is all too often denied to people with profound cognitive disabilities, who are typically assumed to be simpletons with very little in the way of a complex inner life. I do not deny that people with profound cognitive disabilities almost certainly lack some of the complex cognitive capacities that people without such disabilities possess. Despite this, I think it is both false and pernicious to assume that they do not possess a very rich and complex range of likes and aversions, thoughts and perceptions, fears and comforts, that we can at best only guess at in many cases. Not all of these facts about a person's inner life will be easy to discern; some may be expressed very little, or in ways that circumvent even our best efforts to understand.

Silvers and Frances might object that on their view there is no reason a prosthetic reasoner cannot declare that in some situations they are unable to execute all of the reasoning for the subject. To make sense of this claim we would need to hear much more from them as to what grounds the prosthesis would have for making this claim, grounds that do not cast doubt on the whole idea of prosthetic reasoning. It cannot simply be on the basis that the subject does not express any likes or aversions on the matter at hand. I have already stated that a subject who *does* express likes and aversions toward some matter does not need a prosthetic reasoner, just someone who knows him well, including his mode of expressing his will, and who is committed to ensuring his wishes determine what happens to him. Nor will a simple process of extrapolating from expressed preferences be sufficient to answer many pressing questions we have about the person's preferences or values with respect to the complex matters that frequently arise in the lives of people with profound cognitive disabilities. Prosthetic reasoning—executing the reasoning for another with respect to her conception of the good—only seems to have a unique role to play where the subject appears unable to have preferences or values on the matter at hand or is unable to communicate them. And where this is so, it remains mysterious on what grounds a prosthesis may claim serious limits to their ability to execute the reasoning of another without casting the whole notion of prosthetic reasoning into serious doubt.

One can speculate whether an assumption is being made that the complex mental functions of a person with profound cognitive disability can be executed by a diligent trustee because their inner life is at least to some extent a simulacrum of the inner life of the trustee. Think of a more familiar type of behavior, that of a person who believes she can execute and interpret her dog's cognitive processes because she assumes that to a large extent they match her own inner life. This is a very human-centric approach to how we might think of discerning the wills of animals, and it would certainly seem deeply human-centric to suppose we could execute their reasoning processes on their behalf. Not for a moment do I suppose the inner lives of people with profound cognitive disabilities are like those of dogs: they surely would not be, given their human embodiment and active participation in distinctly human practices and human forms of life. Nevertheless, one can query whether there is an "ableist-centric" approach to decision-making embedded within the idea of prosthetic reasoning—namely, that one can access the mind of another person and execute all of their reasoning because it is just like one's own mind, more or less. This, I argue, pays far too little heed to the facts of opacity and fails to express respect for the complex and somewhat ineffable inner lives of others, including those with profound cognitive disability.

Silvers and Frances might object that opacity affects surrogate decision-making as much as it does supported decision-making. If respect requires that we acknowledge that others have rich and complex inner lives to which we only have limited access, does this not also affect a surrogate decision-maker in the execution of their role, and limit what they can claim to know about another person's complex conception of the good?

There are certainly limits to what a surrogate decision-maker can know. But in contrast to prosthetic reasoning, there is nothing within the description and ambition of surrogate decision-making itself that necessarily suggests otherwise. The surrogate decision-maker can decide that they are unable to draw a clear determination as to what the subject really wants, or would want, and thereby revert to other ways of making a decision, including by reference to the person's best interests or well-being. Consider the case of Steve (name changed). Steve has a profound intellectual disability and a range of complex physical disabilities. He is also blind. Steve's greatest joy is food and eating. Despite ongoing attempts to engage him in other activities, Steve shows little active interest in anything other than eating. Unfortunately, he is progressively losing the ability to swallow and is frequently experiencing life-threatening choking episodes. His doctor states that he must take nutrition through a PEG. This will likely prevent Steve from choking to death and ensure he receives adequate nutrition but will also deprive him of the one thing we are confident gives him great pleasure. Unlike Silvers and Frances, I do not believe that we can come to know much about Steve's conception of the good in this case, that is, whether he would value ongoing life more than the joy of eating and thus choose the PEG. I think it would also be deeply presumptuous to declare that the reasoning of the prosthesis (whom Steve may not have chosen or have any control over) has arrived at "his" (Steve's) decision. Rather, as a surrogate decision-maker for Steve I would explicitly state something like the following:

> I am unable to draw any clear conclusions as to what Steve wants or values in this case. He does not appear to express anything on the matter, or what he does express does not lend itself to any clear interpretation. Therefore, we must try to work out what is in Steve's best interests, taking into account all those things about Steve that we do know more about, including his preferences and aversions. But we will never be completely sure that what we end up deciding is the right decision for Steve, or what he would have decided himself if he could.

In explaining their decision in this fashion, the surrogate decision-maker explicitly acknowledges the reality and importance of mental opacity and expresses appreciation for the complexity and ineffability of Steve's mental life. In contrast, supported decision-making for people like Steve assumes that a diligent trustee can come to know his will, or simply execute his reasoning for him, which can then be treated as structurally decisive. I have argued that this assumption rests on morally dubious attitudes about the nature of Steve's inner life and of the kind of relationship others may adopt toward it.

4. CONCLUSION

Replacing legal regimes and practices of surrogate decision-making with supported decision-making is a focus of considerable aspiration among disability advocates and legal scholars worldwide. The arguments of this paper add a philosophical and moral dimension to the cautionary concerns that some legal scholars have expressed about extending supported decision-making beyond where it has real value.[55] None of these authors, including me, are calling for a

[55] Kohn, "Legislating Supported Decision-Making"; and Series, "Relationships, Autonomy, and Legal Capacity."

wholesale retention of surrogate decision-making. I accept that most people with cognitive disability would be capable of making most of their own decisions if appropriate supports were provided and initiated early in life. Supported decision-making in these cases not only respects the rights and interests of people with cognitive disabilities, but also expresses appropriate respect for them as agents. Nevertheless, I have argued that in some cases of severe or profound disability, practices of supported decision-making can express disrespect for people with cognitive disabilities in a number of distinct ways. In some cases surrogate, rather than supported, decision-making will express more respectful attitudes toward this highly vulnerable and stigmatized group of people.[56]

Monash University
linda.barclay@monash.edu

References

Anderson, Elizabeth. "What Is the Point of Equality?" *Ethics* 109, no. 2 (January 1999): 287–337.

Anderson, Elizabeth, and Richard Pildes. "Expressive Theories of Law: A General Restatement." *University of Pennsylvania Law Review* 148, no. 5 (2000): 1504–75

Bach, Michael. "Inclusive Citizenship: Refusing the Construction of 'Cognitive Foreigners' in Neo-liberal Times." *Research and Practice in Intellectual and Developmental Disabilities* 4, no. 1 (2017): 4–25.

Baladerian, Nora J. "Sexual Abuse of People with Developmental Disabilities." *Sexuality and Disability* 9, no. 4 (December 1991): 323–35.

Begon, Jessica. "Paternalism." *Analysis* 76, no. 3 (July 2016): 355–73.

Bigby, Christine, Mary Whiteside, and Jacinta Douglas. "Providing Support for Decision Making to Adults with Intellectual Disability: Perspectives of Family Members and Workers in Disability Support Services." *Journal of Intellectual and Developmental Disability* 44, no. 4 (2019): 396–409.

Dworkin, Gerald. "Paternalism." *Stanford Encyclopedia of Philosophy* (Fall 2020). https://plato.stanford.edu/archives/fall2020/entries/paternalism.

[56] I would like to thank two anonymous reviewers from this journal for their very helpful comments. I would also like to thank audiences from the following two workshops in 2022 for very helpful discussion: "Paternalism and Disability," University of Leeds and Durham University, April 26–27; and "Constructing Social Hierarchy," University of Melbourne, July 11–13.

Flynn, Eilionoir, and Anna Arstein-Kerslake. "Legislating Personhood: Realising the Right to Support in Exercising Legal Capacity." *International Journal of Law in Context* 10, no. 1 (March 2014): 82–104.

Frances, Leslie. "Supported Decision-Making: The CRPD, Non-discrimination, and Strategies for Recognizing Persons' Choices about Their Good." *Journal of Philosophy and Disability* 1 (2021): 57–77.

Frances, Leslie, and Anita Silvers. "Liberalism and Individually Scripted Ideas of the Good: Meeting the Challenge of Dependent Agency." *Social Theory and Practice* 33, no. 2 (April 2007): 311–34.

Groll, Daniel. "Paternalism, Respect, and the Will." *Ethics* 122, no. 4 (July 2012): 692–720.

Hojlund, Anne-Sofie Greisen. "What Should Egalitarian Policies Express? The Case of Paternalism." *Journal of Political Philosophy* 29, no. 4 (December 2021): 519–38.

Horner-Johnson, Willi, and Charles E. Drum. "Prevalence of Maltreatment of People with Intellectual Disabilities: A Review of Recently Published Research." *Developmental Disabilities Research Reviews* 12, no. 1 (January/February 2006): 57–69.

Howard, Dana, and David Wendler. "Beyond Instrumental Value: Respecting the Will of Others and Deciding on Their Behalf." In *Oxford Handbook of Philosophy and Disability*, edited by Adam Cureton and David T. Wasserman, 522–40. Oxford: Oxford University Press, 2020.

Jaworska, Agnieszka. "Respecting the Margins of Agency: Alzheimer's Patients and the Capacity to Value." *Philosophy and Public Affairs* 28, no. 2 (Spring 1999): 105–38.

Kohn, Nina. "Legislating Supported Decision-Making." *Harvard Journal on Legislation* 58, no. 2 (2021): 313–56.

Murphy, Kieran, and Eleanor Bantry-White. "Behind Closed Doors: Human Rights in Residential Care for People with an Intellectual Disability in Ireland." *Disability and Society* 36, no. 5 (2021): 750–71.

Office of the Public Advocate (State of Victoria). *Supported Decision-Making in Victoria*. Melbourne, Australia, October 2020. https://www.publicadvocate.vic.gov.au/joomlatools-files/docman-files/general/Supported_Decision_Making_in_Victoria.pdf.

Pedersen, Viki Møller Lyngby. "Respectful Paternalism." *Law and Philosophy* 40, no. 4 (August 2021): 419–42.

Peterson, Andrew, Jason Karlawish, and Emily Largent. "Supported Decision Making with People at the Margins of Autonomy." *American Journal of Bioethics* 21, no. 11 (2021): 4–18.

Scheffler, Samuel. "What Is Egalitarianism?" *Philosophy and Public Affairs* 31, no. 1 (Winter 2003): 5–39.

Series, Lucy. "Relationships, Autonomy, and Legal Capacity: Mental Capacity and Support Paradigms." *International Journal of Law and Psychiatry* 40 (May–June 2015): 80–91.

Shiffrin, Seana. "Paternalism, Unconscionability Doctrine, and Accommodation." *Philosophy and Public Affairs* 29, no. 3 (Summer 2000): 205–50.

Silvers, Anita, and Leslie Pickering Francis. "Thinking about the Good: Reconfiguring Liberal Metaphysics (or Not) for People with Cognitive Disabilities." *Metaphilosophy* 40, nos. 3–4 (July 2009): 475–98.

Troller, Julien, Preeyaporn Srasuebkul, Han Xu, and Sophie Howlett. "Cause of Death and Potentially Avoidable Deaths in Australian Adults with Intellectual Disability Using Retrospective Linked Data." *BMJ Open* 7, no. 2 (2017): 1–9.

United Nations. Convention on the Rights of Persons with Disabilities. 2515 UNTS 3. Adopted December 13, 2006; entered into force May 3, 2008.

United Nations Committee on the Rights of Persons with Disabilities. General Comment No. 1, Article 12: Equal Recognition before the Law, 2014.

Wall, Steven. "Self-Ownership and Paternalism." *Journal of Political Philosophy* 17, no. 4 (2009): 399–417.

Wasserman, David, and Jeff McMahan. "Cognitive Surrogacy, Assisted Participation, and Moral Status." In *Medicine and Social Justice: Essays on the Distribution of Health Care*, 2nd ed., edited by Rosamond Rhodes, Margaret Battin, and Anita Silvers, 325–33. New York: Oxford University Press, 2012.

Wolff, Jonathan. "Fairness, Respect, and the Egalitarian Ethos." *Philosophy and Public Affairs* 27, no. 2 (Spring 1998): 97–122.

Journal of Ethics and Social Philosophy
vol. 27, no. 1 · march 2024

https://doi.org/10.26556/jesp.v27i1.2667

© 2024 Author

FORGIVENESS AND NEGATIVE PARTIALITY

Joshua Brandt

Partiality and forgiveness are both characteristically *personal* dimensions of morality. Forgiveness requires having the relevant standing as victim (or being closely connected to the victim), and reasons of partiality are agent relative, being derived from an agent's relationships or histories of interaction. I argue that an integral connection between these phenomena emerges once an expansive concept of partiality is adopted—one that includes the negative analogue of intrinsically valuable relationships, such as friendship and family. While positive partiality involves the acquisition of special permissions or duties to promote another's interests, relationships of negative partiality involve the acquisition of special permissions or duties to discount interests. I argue that forgiveness should be conceptualized as a way of ending these negative relations.

In relationships of justified partiality, members are closer together in moral space, and justified negative partiality analogously reflects a kind of moral distance (strangers representing a midway point).[1] Forgiveness eliminates the moral distance within a negative relationship by altering the norms that it otherwise grounds.[2] This metaphor is made concrete in the proceeding analysis as the notion of negative partiality is clarified. But why accept this understanding of forgiveness? My approach draws from well-recognized considerations of theoretical adequacy developed in the literature: a theory of forgiveness should fit (and, ideally, explain) the personal nature of forgiveness and the normative significance of forgiveness, and it should distinguish forgiveness from related phenomena (e.g., excusing and justifying). Ideally, a complete theory of forgiveness will also capture nonparadigm cases of forgiveness, such as third-party forgiveness and self-forgiveness. The theory I present has this explanatory power.

[1] This metaphor, first employed by Broad in "Self and Others," is discussed in further detail below.

[2] As discussed below, Bennett and Warmke both endorse versions of the view that forgiveness alters moral norms. The contrast between our views will be explored in the second half of this paper.

I proceed by critiquing the prevailing view that forgiveness is exclusively a descriptive phenomenon (i.e., either a psychological process or a behavioral pattern). This discussion motivates a normative standard of theoretical adequacy for forgiveness, a condition that the framework of negative partiality fits and explains. The latter sections of the paper compare my view with related positions that understand forgiveness as a normative power, arguing that conceptualizing forgiveness in terms of negative partiality more plausibly delimits the scope of this power and avoids counterexamples faced by competing views. I conclude by considering some important ways in which descriptive and normative accounts of forgiveness may be related.

1. Theoretical Adequacy: Initial Remarks

My view shares some common (though not entirely uncontroversial) propositions concerning forgiveness with existing accounts. First, forgiveness is a response to a wrongdoing: a homeowner might "forgive" the neighborhood children for breaking their window, but the possibility of forgiving is undercut if, in fact, the window was broken by a stick carried by a gust of wind. To forgive simultaneously construes an act *as* wrong and, in some sense, extinguishes the wrongdoing; a permissible act cannot be extinguished in the relevant sense. Second, forgiveness responds to the blameworthy. A driver who rear-ends another vehicle for a nonculpable reason (e.g., the driver suffered a heart attack) could be excused, but not forgiven. Third, I assume that forgiveness is a personal response to culpable wrongdoing. Only the victim (or sometimes a person closely connected to the victim) has standing to forgive. Presuming the neighborhood kids *did* break the window, the victimized homeowner (and not simply any person across town) can forgive. Relatedly, forgiveness is the prerogative of the individual who has this special standing: forgiveness *may* or *may not* be granted, but the decision is up to the victim with the relevant standing.[3] Thus, the prerogative to forgive and the standing to forgive represent two distinct senses in which forgiveness is "up to" the victim.

In what follows, the conditions of theoretical adequacy are further developed by examining existing accounts in closer detail. I argue, following Warmke and Bennett, that forgiveness is normatively significant.[4]

[3] Recognizing this prerogative presumes the *general* permissibility of forgiveness (there may be exceptions as discussed below).

[4] Bennett, "The Alteration Thesis;" Warmke, "The Normative Significance of Forgiveness."

2. The Normative Significance of Forgiveness

The prevailing approach to forgiveness represents this phenomenon as an emotional process. These affective accounts typically cite Bishop Butler's idea that forgiveness is the "forswearing of resentment" as their point of departure, the forswearing of resentment understood as a descriptive psychological process.[5] While forswearing amounts to ridding oneself of negative feelings directed toward the perpetrator of a wrongdoing, proponents of the affective view have plausibly argued that *merely* overcoming resentment is insufficient. One might simply *forget* a past wrongdoing and on this basis overcome resentment—but forgetting is not forgiving. One might also overcome resentment through behavior modification therapy, but this, too, is not forgiveness.[6] More sophisticated affective accounts add conditions that explain *how* resentment must be overcome. Murphy's classical statement of this thought is that resentment must be overcome for moral reasons to qualify as forgiveness (e.g., because a perpetrator has apologized).[7] Instead, it could be argued that forgiveness involves *seeing* the perpetrator as a decent person (one worthy of reconciliation).[8] Alternatively, it could be argued that forgiveness involves a reevaluation of a person's character in a way that excises the particular wrongdoing for the purpose of assessment.[9]

In what follows, I draw attention to what is left out by a *solely* descriptive account of forgiveness.[10] This critique requires the further observation that forgiveness changes the "moral standing" between the victim and the perpetrator of a wrong—i.e., their relationship departs from the baseline relation of equality that typically holds between persons. Here, I rely on the assumption that by *wronging* another, the perpetrator is, in some sense, a less worthy person (at

[5] Butler himself may not have endorsed this psychological interpretation of forgiveness. See Newberry, "Joseph Butler on Forgiveness."

[6] For relevant discussions, see Murphy, *Getting Even*; Hieronymi, "Articulating an Uncompromising Forgiveness," 529–55; Allais, "Wiping the Slate Clean."

[7] Murphy, *Getting Even.*

[8] Murphy and Hampton, *Forgiveness and Mercy*.

[9] Allais, "Wiping the Slate Clean," 33–68.

[10] In drawing attention to the idea that there are practices left out by solely descriptive accounts of forgiveness, I do not intend to draw the further conclusion that we must outright reject descriptive accounts of forgiveness. Descriptive theories have offered insight into the psychological dimension of forgiveness but, as argued in this section, do not fully capture a normative dimension of this practice. Insofar as the psychological and normative use cases of the concept of "forgiveness" come apart, we may need to be pluralistic with respect to the concept of forgiveness. This question is taken up further in the concluding section of the paper.

least from the vantage point of the victim). For this reason, the perpetrator has a good reason to want forgiveness, and there should be something that changes from the moral point of view once the victim has forgiven. My understanding of standing is clarified in the next section; however, to offer an example from a different context, we might say that by violating the law, a criminal has lowered their standing in the community—i.e., the community can punish them or owes them less. To imagine how an affective view would explain this idea, suppose that a victim forswears resentment because the perpetrator has apologized, and the victim now sees the perpetrator in a better light. This change must somehow explain why the moral standing of the perpetrator improves. However, if we appeal to the intrinsic features of forgiveness within the affective view, the explanation seems unsatisfying. After all, merely *being* resented does not make a person have lower moral standing with respect to another. One can resent a person, even if they have done nothing wrong; giving up this negative attitude cannot improve standing as there was no unequal standing to begin with.

The natural defense of the affective view that maintains an intrinsic connection between forgiveness and moral standing is to note the requirement that forgiveness responds to wrongdoing. Given this fact, the affective theorist could argue that it is not resentment, in general, that lowers a person's moral standing, but only resentment when based on a wrongdoing; in other words, they could argue that resentment does not have the power to alter a person's moral standing when there is *no reason* for resentment, but it can when appropriately grounded.

I do not believe this addendum resolves the issue. To illustrate why, consider the following case.

> *Wavering Wally*: Wavering Wally was wronged by a former undergraduate colleague, Molly, but he has long since forgiven her for the past misdeeds. It has been years since Molly plagiarized Wally's term paper and their eventual reconciliation. However, Wally recently finds himself wavering in these feelings about the past: he experiences bouts of resentment followed by the dissolution of such feelings. After some time, the back-and-forth process shows no sign of letting up: currently, he resents her.

What should the affective account of forgiveness say about the case of Wavering Wally? Two possibilities suggest themselves, each seemingly problematic. It might be claimed that Wally really has forgiven Molly. This analysis is consistent with the fact that a long period of time passed during which Wally had given up his resentment. However, it is unclear how the affective view can make this claim. After all, Wally has not fully overcome his resentment: he still presently resents her (on the basis of a wrongdoing). On the affective view, it further seems that Wally's wavering emotional state must cause Molly's moral standing to be lowered once again. This implication follows from the fact that it is *resentment* (based on wrongdoing) that is responsible for lowering Molly's standing. This implication seems problematic: Molly, who previously occupied a higher standing, has now been lowered merely in virtue of a change in Wally's emotional state. Why should forgiveness, on this view, be morally significant? Even declarations of the form "I forgive you" are subject to a revision in feelings.

On the other hand, one might argue that Wally never forgave Molly, and it is for this reason that her standing is still low when he later becomes resentful. However, this interpretation also seems implausible: if Wally has not forgiven Molly, it explains his current resentment, but it leaves the large gap of time when it *seemed* as if he had forgiven her incorrectly described. On this view, all apparent forgiveness was undermined by the fact that he later came to resent her (suppose it has been twenty years!). Should we say, following Aristotle on happiness, that a person cannot truly have forgiven until they are dead? On this interpretation, a declaration of forgiveness will, again, retroactively not count for anything given a revision in attitude. After all, the revision shows that there had been no forgiveness.[11]

An alternative approach for affective accounts is to draw an indirect link between forgiveness and standing.[12] For example, it could be argued that forgiveness impacts a relationship, and it is the relationship that ultimately determines standing. Consider a case where resentment is a barrier to a friendship—the victimized party may wish to reconcile with the perpetrator but find themselves partly incapable of living up to the norms of the prior relationship (e.g., they may be unable to celebrate the success of the perpetrator). In a more extreme scenario, negative attitudes may completely undermine the relationship. A proponent of the affective model could plausibly argue that overcoming resentment (e.g., by reevaluating how a wrongdoing figures in the assessment

[11] We might instead say that Wally *had* forgiven her, but no longer forgives. But on this reading, Wally's emotional wavering is still capable of altering Molly's moral standing. In the past, when Wally had forgiven Molly, her standing was higher, but now that he no longer forgives, her standing is lower.

Warmke draws from a related case to show that forgiveness has the normative effect of *obliging* the victim to treat the perpetrator according to altered norms (a conclusion I likewise endorse). I believe the modified scenario described above shows something further: purely affective accounts of forgiveness are unable to provide a complete theory of forgiveness insofar as they cannot diagnose cases of "wavering" emotional responses that take place over long periods of time.

[12] This defense on behalf of the affective view was offered by an anonymous referee.

of a person's character) contributes to a normalization of a relationship, which in turn improves moral standing.

While affective shifts plausibly contribute to reconciliation, there remain difficulties with linking moral standing to the effects that forgiveness has on a relationship. A problematic implication of this view is that forgiveness remains contingently connected to moral standing. Suppose a (victimized) friend overcomes resentment yet makes the calculated decision to dissolve the friendship. We cannot appeal to the relational effects to explain any shift in moral standing (there will be no such effects). A second difficulty with an appeal to the significance of relational shifts is that forgiveness may take place outside of the context of a morally significant relationship. The victim of a pyramid scheme may come to forgive the perpetrator yet never have been in a morally significant relationship with them to begin with. Where there is no preexisting relationship, it is difficult to cite relational effects to explain an improvement in moral standing.[13]

The above considerations suggest that forgiveness does not merely track reactive attitudes. More generally, for any stipulated descriptive criteria of forgiveness, we can imagine a scenario where forgiveness has not "really" taken place because the victim alters their attitude or behavior; insofar as the perpetrator's standing is subject to revision, there is difficulty capturing the normative significance of forgiveness. The dilemma generated by this analysis motivates my endorsement of views that draw a connection between intrinsically normatively significant acts, such as promises and forgiveness. On these views, forgiveness should be understood as an act that alters the moral standing between parties by giving up certain rights that were previously possessed by the victim of the wrongdoing.[14] If Molly has a legitimate complaint against

[13] A potential response is that where there is no preexisting relationship, people are, by default, *open* to relationships with each other. The affective view may claim that resentment is a barrier to this openness, a situation that dissolves with forgiveness. While this model fits some cases, I worry that it will not capture the full scope of possibilities. Two people who share a workspace may have no desire to form a relationship—i.e., they may not be open to friendship from the outset of knowing each other. While a default attitude of openness might be common (or, perhaps, a virtue), a victim who started from a place of being closed off will have no way to improve the relationship with the perpetrator on the affective model. There may, in some cases, be good reasons for individuals to be closed to a relationship with each other (despite the fact that nothing wrong has occurred)—perhaps they know that they have nothing in common or simply find each other annoying. When forgiveness between such "incompatibles" occurs, I would be inclined to say that the relationship has improved, but only through changes in negative partiality (as outlined below).

[14] Bennett terms this view the "alteration" thesis, and Warmke likewise endorses a version of it.

Wally *because* he resents her, we should conclude that he gave up the right to resent in his initial performative act of forgiveness.

3. Moral Alterations and Negative Partiality

I argue that the moral alteration brought about by forgiveness should be understood in terms of partiality. To establish this view, I provide a sketch of how (positive) partiality should be understood, subsequently employing the negative analogue of this relation as a way of understanding forgiveness. Justified positive partiality involves a departure from the ordinary requirements of impartial morality: friends and family have special duties or permissions to promote each other's interests. For example, on most plausible moral theories, there is a *prima facie* duty of beneficence—i.e., the beneficial effects of an action provide a reason to perform the action; partiality can be thought of as strengthening this duty. In this form of partiality, all else being equal, there is a duty to promote the well-being of one's intimate rather than the equal well-being of a stranger. C. D. Broad pictured our moral relations as a series of concentric circles in moral space: an "innermost circle" of individuals representing those to whom we owe the most, with sequential circles representing decreasing degrees of intensity in our obligations.[15] On Broad's picture, the outermost circle represents strangers, to whom we have the weakest obligations.

Broad's spatial metaphor is an attractive way to capture the idea that our duties have varying weights depending on the significance of a moral relation, but this picture should be expanded to include further variations of partiality. First, our partiality can be conceptualized in terms of *permissions* in addition to duties. If we assume a background normative theory that recognizes a *prima facie* duty of beneficence, partiality can be understood as a special permission that *allows* an individual to prefer the well-being of their intimates over strangers. On this view, an individual has a prerogative to promote the lesser well-being of their intimate over the greater well-being of a stranger.[16] This form of partiality never *requires* that one prefers the interests of one's intimate. Therefore, underived special permissions and obligations represent two distinct dimensions of partiality (in contrast with the single dimension suggested by Broad).[17]

The second major modification to Broad's spatial metaphor targets the idea that the outermost circle is occupied by strangers. We conceptualize negative

[15] Broad, "Self and Others."

[16] Scheffler was an early proponent of this view. See Scheffler, *The Rejection of Consequentialism*.

[17] For an in-depth discussion of underived permissions, see Hurka and Schubert "Permissions to Do Less Than the Best."

partiality by supposing that moral distance goes beyond strangers. This relation can be explained in terms of negative analogues of the changes described above. Just as positive partiality can involve the baseline duty of beneficence strengthening, negative partiality can be thought of as this duty weakening. The negative relation can be understood in yet a stronger fashion if the initial duty of beneficence is "inverted"—i.e., the fact that an action benefits a negatively related individual counts as a reason *against* performing the action. Importantly, the weakened *prima facie* duty of beneficence does *not* imply a duty to harm, but rather a duty to prefer the lesser well-being of a stranger over the greater well-being of the individual who stands in a negative relation (the inverted duty likewise implies no duty to harm).[18] And just as with the positive relation, we can conceptualize negative partiality in terms of a special permission to discount the interests of a particular person.

With a sketch of how negative partiality can be characterized, I return to the case of an individual victimized by plagiarism. I propose that forgiveness in this context should be understood to extinguish the relevant manifestation of negative partiality justified by the historical relationship between these parties. After the incidence of plagiarism, Molly and Wally might encounter each other in various contexts around campus. Wally plausibly has, in at least some cases, a special basis for discounting Molly's well-being, thereby varying from the impartial requirements of morality. Suppose, for example, that Wally can assign his rent-controlled lease: it strikes me that he may choose, based on his past victimization, to ignore Molly's application in favor of a stranger with whom he has no connection. In this case, negative partiality plausibly involves a permission to prefer the lesser good of the stranger over the negatively related individual (e.g., supposing that Molly stood to gain *more* from living in the apartment, Wally *still* seems permitted to prefer the well-being of the stranger). I would argue further that, within reason, Wally may discount well-being in *nondistributive* cases. Suppose, for example, that study groups regularly meet on campus and that Wally overhears a now-reforming Molly expressing a desperate need to somehow raise her grade. For any other student, it would strike me as a basic fulfillment of the duty of beneficence to offer information about the group, but Wally seems permitted to avoid volunteering this information. To be precise, I would not describe Wally as being *required* to avoid assisting Molly in the above ways: it is his prerogative to discount or not.[19]

[18] Of course, negative partiality *can* be conceptualized as a permission or duty to harm, but these normative changes are not a *necessary* characterization of the relation.

[19] A question at this stage regards the precise *scope* of the permission of negative partiality. For example, it does not seem like the victim of plagiarism can prefer to save a stranger over the perpetrator in a rescue scenario. My view is that where increasingly significant welfare

There are ways to be skeptical about negative partiality: some theories of desert hold that vice warrants impartial censure and proportionate suffering. It could plausibly be argued, for example, that insofar as Molly has victimized Wally, *everyone* has reasons to discount Molly's interests. Such a position challenges the idea that Wally has *special* reasons to discount Molly's well-being and, therefore, the broader thesis that Wally has acquired special standing to forgive. In response, it may be helpful to consider how analogous concerns could be raised with respect to the justification of *positive* partiality. For example, we often see a great friend as simultaneously a great person. But despite the legitimate sense in which a good friend (and a correspondingly good person) may be the more deserving recipient of benefit, friendship is plausibly characterized as a relationship of justified partiality. There is an explanation for this analysis: while good friends may have underlying virtuous dispositions, friends still benefit each other in ways that go above and *beyond* what is required by impartial desert. Suppose that *A* is friends with *B* and *Y* is friends with *Z*, and each individual is aptly characterized as a great friend. While each is (by stipulation) an equally deserving person, it still seems that *A* may prioritize the interests of *B* (over *Z* or *Y*), and *Z* may prioritize the interests of *Y* (over *A* or *B*).

Returning to the case of Molly and Wally, I would argue that even if Molly warrants less from the impartial point of view (i.e., we suppose that everyone may discount her interests), Wally still has special reasons for discounting. To test this hypothesis, we could imagine an idealized scenario that mirrors the pairs of friendships described above. Suppose, for example, that Wally is assigning his lease and must choose between either Molly or the perpetrator of a wrongdoing similar in degree (e.g., some *other* plagiarizer). Must he choose to distribute the benefit impartially, or can he prefer the stranger? Intuitively, I would argue that the stranger may be preferred. A more pedestrian example arises in cases of infidelity: while the victim of a breach of trust within a

is at stake, a more serious wrongdoing is required to justify corresponding negative partiality (i.e., only a serious wrongdoing could possibly justify discounting a person's welfare in a life-and-death scenario). Analogously, while union members plausibly have duties of *positive* partiality to each other, being members of the same union does not obviously allow for preferential treatment in a rescue scenario. See Dworkin, *Law's Empire*; Stroud, "Permissible Partiality, Projects, and Plural Agency"; and Davis, "Scope Restrictions, National Partiality, and War." Similarly, scope restrictions on partiality apply in the context where an individual explicitly undertakes a role that requires distributing a good in an impartial fashion. For example, a physician may not seek to manipulate an organ-donation list to prioritize their loved one, nor may a judge seek to apply a reduced sentence for the sake of an old friendship. Likewise, the scope of negative partiality should be restricted to exclude encounters that take place where the victim undertakes a role that explicitly requires the application of impartial rules (e.g., where the victim is judging an athletic competition). See Cottingham, "Partiality, Favouritism and Morality," for a classical discussion.

relationship may have strong reasons for moral distancing (perhaps obligations for this response), it seems implausible (and overly punitive) to extend reasons of identical strength to all prospective romantic partners.

Further considerations support interpreting the case as one involving negative partiality, rather than a straightforward application of justice. While plagiarism warrants punishment, the correct office for fairly distributing it is most plausibly a university body; the happenstance interactions that transpire between victim and perpetrator hardly count as impartial justice. Moreover, we can stipulate for dialectical purposes that Molly's conduct is mitigated by the overall assessment of her character. Perhaps she has previously been an upstanding member of the university community, only driven to plagiarism by overstretching herself in service to student governance. While such mitigating factors could plausibly undermine third-party reasons for negative responses (i.e., how the general community of students should respond to Molly), they nonetheless seem unable to undermine the victim's special standing to discount interests.[20]

While the above motivates the intuitive case for negative partiality, it may be further asked *why* such reasons for action are generated by a history of victimization. A full defense of the grounding of negative partiality is beyond the scope of this paper, but I hope to gesture at how the issue can be approached. Discussions of partiality have largely focused on how preferential treatment among friends, family, and other special relationships can be justified in light of the apparent impartial demands of traditional moral systems. Given, for example, the apparent equal significance of interests "from the point of view of the universe," impartialists ask why intimates may attach greater weight to each other's interests. On my view, understanding how partialists have replied to this challenge can inform the analogous relation of negative partiality, wherein less weight is attached to the interests of a perpetrator of a wrong.

[20] *Mutatis mutandis*, the reasons of negative partiality possessed by Wally do not reduce to a collectively held right possessed by the university community to discount Molly's interests. While it is true that all members of the community may have a special reason to stand up against plagiarism (because they are members of the community impacted by the wrongdoing), Wally has stronger reasons for such responses. We can observe this difference in the two scenarios described above: Wally may discount the interests of the individual who plagiarized *his* work over some other individual who plagiarized, and Wally's reasons for discounting Molly's interests are not undermined by the fact that Molly has contributed to the university community (by contrast, the community's reasons for negative responses may significantly be diminished by such factors, even if not eliminated). I thank an anonymous reviewer for raising this objection.

For a more developed discussion of negative partiality, see Brandt, "Negative Partiality" and Lange, "Other-Sacrificing Options."

Among the most prominent approaches to justifying positive partiality is appeal to an agent's projects (e.g., Stroud) and appeal to the value of special relationships (e.g., Scheffler).[21] Stroud argues that a special permission to pursue one's projects is needed to push back against the excessive demands of consequentialist obligations, which are at odds with the nature of human agency. Since special relationships are a class of projects that *require* partiality, the special duties of partiality are justified indirectly through an agent's projects. Those who justify partiality by appeal to the *value* of such relationships argue that morality must make room for the intrinsic value of friendship, family, etc., and that such relationships can only exist if partiality itself is permitted. Again, partiality is indirectly justified, in this case by appeal to its role in bringing about intrinsic value.

If we are to move beyond appeal to the intuitive justification of negative partiality, we might deploy known justifications of partiality to the negative sphere. A victim of wrongdoing may very well transform their relationship with the perpetrator into a project of personal significance. Mirroring positive relationships, which require partiality, the project that a victim undertakes plausibly could require discounting the interests of the perpetrator (they may likewise take up the related project of standing up for themselves, which could also involve discounting the interests of the perpetrator). There will, of course, be outstanding questions for this approach; for example, where will projects of negative partiality be themselves justified? Moving to the second approach, while negative relationships are not traditional candidates for what people value, a victim can plausibly value their resistance and opposition to the perpetrator of a wrong. By resisting the perpetrator, the victim enters a kind of relationship worth valuing, one that involves negative partiality. We can, therefore, construct a mirror of another prominent grounding approach to positive partiality. Insofar as morality must make room for relationships that can reasonably be valued, and such relationships involve negative partiality, negative partiality will be justified.

Both approaches described above give rise to further questions and challenges, but general strategies for grounding negative partiality can draw from known resources. In arguing that forgiveness should be grounded by negative partiality, I do not take a particular stance with respect to *which* approach we should endorse, but I take it that a range of options is open and compatible with the analysis that follows.

A straightforward explanation of what forgiveness accomplishes and why it is normatively significant follows from the above framework. Morality recognizes victims by empowering them to discount the interests of the perpetrator.

[21] Scheffler, *The Rejection of Consequentialism;* Stroud, "Permissible Partiality, Projects, and Plural Agency."

A victim who forgives surrenders this power to discount, and it is for this reason that Molly is entitled to complain when Wally "wavers" in his forgiveness. To explain the varieties of forgiveness, we should consider the further normative question of whether the victim needs a *reason* to forgive. The present framework addresses questions of justification (or reasons for forgiveness) by considering the specific nature of the normative change that results from a wrongdoing. As noted above, the victim of plagiarism plausibly acquires *a special permission* to discount the perpetrator's well-being—i.e., a right to discount or not to discount. Insofar as the victim possesses a right of this kind, it may be abandoned at will. An analogy to promising seems fitting here. An individual who promises to ϕ surrenders their right to $\sim\phi$, and *insofar* as they are permitted to ϕ from the outset, they do not require a *reason* to promise to ϕ. A promise to meet someone for lunch thereby surrenders the right to do otherwise, but this promise does not require justification if the lunchtime meeting was permissible from the outset; likewise, forgiveness surrenders the right to discount a person's well-being, and given a standing permission of negative partiality, no special reason is needed to give up the right.[22]

It is nonetheless straightforwardly compatible with this framework that there can be good reasons to forgive. A person who, for example, apologizes or provides compensation may be worthy of forgiveness—however, these reasons are simply not required for and do not compel forgiveness. Consider again the analogy to promising. There can be better or worse reasons to promise to help your friend move: perhaps they have helped you in the past, or asked nicely, or desperately needed the help. Despite these good reasons to promise, the promise does not require them to gain normative force.

The victim-perpetrator relationship arising in the plagiarism case represents a paradigm instance of forgiveness that captures the prerogative of forgiveness. But the normative assessment of forgiveness might not be limited to this kind. Other varieties of negative partiality explain how we assess other cases of forgiveness. Consider, for example, the idea that forgiveness ought, in some cases, to be *conditioned*—i.e., it would be either impermissible or *impossible* to forgive an individual until certain conditions are met (repentance, apologies, reparations, etc.).[23] Conditional forgiveness is captured by the idea that the

[22] For a contrasting view, see Milam, "Reasons to Forgive." Milam argues that "accounts of forgiveness as cancelling a moral debt" (such as the views of Warmke, Bennett, and myself) do not distinguish between deciding not to blame and forgiving (246). After all, one may cancel a debt by deciding the perpetrator was not blameworthy *or* forgiving, and so it will be unclear how the two are distinct on the debt-cancellation view. I respond to this specific concern below.

[23] For a defense of the view that forgiveness ought to be conditional, see Haber, *Forgiveness*.

victim is *obliged* to engage in distancing responses—i.e., they have duties of negative partiality conditional upon the performance of certain acts by the perpetrator (or certain circumstances being present, such as the suffering of the perpetrator). My purpose here is not to take a substantive position on the question of whether forgiveness is sometimes conditional or not, but to show that the framework of negative partiality has an explanation of this possibility. Those attracted to conditional forgiveness plausibly endorse the corresponding idea that victims are (conditionally) prohibited from engaging in certain forms of beneficence (e.g., in virtue of self-respect). Consider, for example, whether the former partner of an unrepentant philanderer charged with operating an online romance scam should contribute to their ex-partner's legal defense fund. In such a case, many people would be inclined to say that the victim not only *may* refrain from offering assistance but that they ought to refrain (perhaps reflected by the imperative to "stand up for yourself"). Clearly, however, such a prohibition cannot be an *impartial* prohibition as considerations of impartial justice generally *support* access to qualified legal assistance.

Negative partiality, therefore, has the resources to explain both cases of conditional and unconditional forgiveness, the former grounded by the prerogative of negative partiality and the latter being captured by a duty of negative partiality. Once the relevant conditions have been satisfied, the victim shifts from having a duty of negative partiality to possessing the prerogative, and the paradigm model of forgiveness will apply. The explanation might proceed as follows: insofar as the perpetrator has apologized, the victim will no longer be acting in a way that compromises their self-respect when they choose to no longer discount the interests of the perpetrator. The victim is nonetheless still *entitled* to discount the interests of the perpetrator. Importantly, this picture preserves the sense in which the victim has the prerogative to forgive in both conditional and unconditional cases of forgiveness: even when the victim is prohibited from forgiveness, it will ultimately be up to them (and not anyone else) to forgive.

Before moving to the theoretical virtues of this account of forgiveness, a final question concerns the application of this theory within friendship, family, and other relationships of positive partiality. Imagine, for example, a breach of trust in the context of an otherwise long-standing and great friendship (suppose that Andy violates the confidence of Lesley by carelessly revealing sensitive information about her marriage). It is implausible to suppose that such a breach *necessarily* justifies treating Andy as an individual who is owed less than a stranger—i.e., it is possible for a friendship to withstand such a wrong. We might ask the following: if forgiveness involves surrendering the right to *negative* partiality, how is forgiveness possible when a relationship ultimately remains a case of partiality? On my analysis, if friendship (and thus partiality) withstands a wrongdoing, two factors are present in the relationship: a wrongdoing that generates a permission to attach less weight to the interests of the friend than would otherwise be permitted and a history (e.g., a shared history of mutual beneficence and intimacy) that grounds a stronger *prima facie* duty of partiality. On this picture, wrongdoing in the context of an intimate relationship does not transform the relationship into enmity but still generates "moral distance"—i.e., an agent-relative reason that justifies discounting the interests of the perpetrator.

While a relationship may remain a case of (overall) justified partiality, the *prima facie* permission generated by the wrongdoing still alters the norms of the relationship. In particular, the victim will now be allowed (but not required) to discount the well-being of the perpetrator when it would have otherwise not been permissible (this relation might be termed "relative negative" partiality since the moral distance is relative to the higher baseline of beneficence in the relationship). Suppose, for example, that shortly after the breach of trust, Andy asks Lesley for assistance with his *own* relationship difficulties. While offering a patient and sympathetic ear might have otherwise been the unquestioned requirement of their friendship, it is natural to see how the breach in trust allows Lesley to be distant; of course, the distance created by a wrongdoing may be more subtle: Lesley may simply now have a legitimate basis for being less responsive to the overall maintenance of the friendship (e.g., by withdrawing from shared projects and activities). Thus, while Lesley may still have "net" duties of partiality to Andy (e.g., she would be present for him in ways that she would not be for others), she simultaneously has a special basis for discounting his interests. In forgiving, Lesley surrenders this claim to discounting and thus eliminates the moral distance present in the relationship.

In sum, forgiveness within partial relationships is continuous with forgiveness in the context of other interpersonal relationships. In each case, the victim acquires a special claim against the perpetrator to discount their well-being and surrenders this claim in forgiving. The central difference between these cases is where the baseline duty of beneficence is set prior to the wrongdoing. Where individuals have no relationship, forgiveness surrenders the right to negative partiality (understood in the strict sense outlined at the outset of the paper), and where there is a positive relationship, forgiveness surrenders the rights to "relative" negative partiality (i.e., the right to ignore the stronger duty of beneficence that would otherwise be present in the relationship).

4. EXPLANATORY ADEQUACY

Having illustrated how negative partiality can model core cases of forgiveness, I turn to the theoretical virtues of this view. Consider, first, the well-established notion that forgiveness is a *personal* relation—i.e., only certain individuals have the standing to forgive. The view that forgiveness involves surrendering the right to negative partiality captures this idea. After all, justified partiality is conceptualized in terms of agent-relative reasons for action—i.e., only individuals who stand in the relevant special relationship have reasons of partiality. Many accounts of forgiveness *stipulate* that only the victim of a wrongdoing may forgive, but the present view *explains* this fact: it is because the victim acquires the prerogative of negative partiality that they may forgive—individuals without this right cannot surrender it.

The present view also offers a natural explanation for why forgiveness responds to a blameworthy wrongdoing and is thereby distinct from acts of "justification" or "excusing." The most plausible ground of negative partiality is culpable wrongdoing. By contrast, to discount someone's interests for poor reasons (e.g., because of a person's taste in music) does not thereby reflect genuine moral distance; one may *act* as if there is moral distance, but poor taste in music does not justify negative partiality. If a blameworthy wrong is a necessary condition of negative partiality, then we can explain why the neighbor who suffers from a broken window carried by a gust of wind is unable to forgive the neighborhood kids. Forgiveness surrenders the right to negative partiality, but no such right is present in this case. Likewise, an individual who overcomes a negative affective/behavioral disposition by recognizing that they were never wronged is not thereby surrendering any special rights of negative partiality.[24] It is for this reason that we say they are "justifying" the act, rather than forgiving. Similarly, excusing involves recognizing that an act *would* have generated reasons of negative partiality (but for some special consideration) and is, therefore, also distinct from forgiving.

The present account of the grounds of negative partiality could, of course, be questioned. One might reasonably argue, for example, that a permissibly inflicted harm generates reasons of negative partiality. Consider a family-operated flower shop that has been run for generations, only to face stiff competition from a new entrant; over the course of a few years, a price war ensues, and the entrant ultimately prevails. Some might be sympathetic to the idea that despite the permissibility of the new store's conduct (assuming fair competition), the family has a legitimate basis for resentment and corresponding acts of negative

[24] This addresses Milam's concern that the debt-cancellation model of forgiveness does not distinguish between ceasing to blame and forgiving. The essence of my response is that one cannot successfully surrender a right (i.e., genuinely forgive) if by the very nature of "surrendering" the right, they deny having had the right in question. Insofar as ceasing to blame involves recognizing that an act was not culpably wrong, the person who ceases to blame makes no claim to having a right to negative partiality (or giving it up).

partiality. Perhaps they retaliate by deploying their vehicles in the most convenient loading zones of the entrant's storefront or lobby in support of Walmart's effort to have the area rezoned for an even bigger commercial enterprise. While I am unsympathetic toward the idea that these responses are permissible, let us grant a hypothetical interlocutor the case. Inasmuch as one endorses reasons for negative partiality in this case, I believe our interlocutor would *likewise* be inclined to endorse the idea that there *is* something to forgive in this scenario. If the now-impoverished family has a right to the aforementioned acts, they can presumably surrender the right in question through another act that would be characterized as forgiveness. Notably, this interpretation would force us to revise a widely held condition of forgiveness—i.e., forgiveness responds to a wrong. This brief dialectic supports the general idea that our understanding of forgiveness is *ultimately* informed by negative partiality: insofar as we expand the grounds of negative partiality, we likewise expand the cases where forgiveness is present. I believe this connection is theoretically significant. It is striking that the background conditions which render forgiveness possible are aligned with the grounds of negative partiality. A plausible inference is that forgiveness is to be explained in terms of negative partiality.

Before moving to nonparadigm instances of forgiveness (i.e., self-forgiveness and third-party forgiveness), there are two final explanatory considerations to consider. First, I propose that there is an attractive *disconnect* between the pursuit of justice and forgiveness on the view presently defended.[25] For example, it will be straightforwardly consistent with forgiveness to testify against the perpetrator of a crime, pursue them actively in court, and publicly affirm the appropriateness of punishment. These actions are compatible with forgiveness because in surrendering the right to negative partiality, the victim makes no statement regarding the appropriateness of impartial punishment. Of course, a judge may look to the fact that the victim has forgiven as a way of evaluating whether the perpetrator has sincerely felt guilt, made amends, and so on. Though, even this connection between forgiveness and justice must be qualified. An act of *unconditional* forgiveness will not be evidence of reformation in the perpetrator. By contrast, it *does* make sense to look at the reasons motivating forgiveness as evidence of the mitigating factors relevant to punishment.

And finally, the present view explains the sense in which forgiveness reestablishes moral equality between victim and perpetrator. Partiality represents a paradigmatic departure from equality. For this reason, positive partiality has historically requested justification: we must explain why parents, friends, and

[25] Allais articulates important concerns with several views that fail to disconnect forgiveness from justice. See Allais, "Wiping the Slate Clean"

other associates can treat the interests of their intimates with greater urgency than those of strangers. In the case of *negative* partiality, the justificatory ground begins with wrongdoing, and this explanation seems in keeping with the widely held view that the perpetrator of a wrong occupies a lower moral standing, albeit one that can be improved by forgiveness. The metaphorical language of "higher" and "lower" standing is made concrete by the idea that the victim possesses a special right to discount the interests of the perpetrator, and it improves the perpetrator's standing by surrendering the right in question.

5. Third-Party Forgiveness and Self-Forgiveness

While it is widely acknowledged that forgiveness is not an *impersonal* phenomenon (i.e., a restricted class of individuals has standing to forgive), it has been plausibly argued that this standing should be expanded to include close relations of the victim. One interpretation of such forgiveness is that it is purely grounded on the indirect victimization of these relations themselves. For example, relatives of a victim may be harmed insofar as they are distressed by the suffering of their loved ones. Alternatively, some theories of well-being hold that individual welfare is intrinsically impacted by the happiness of one's relata; it could be argued that the lives of parents go well, in part, to the extent that their children are happy. On this view, wrongfully harming a child indirectly victimizes the parents who then acquire the standing to forgive.[26]

If third parties are, in fact, victims, then forgiveness will apply in the paradigmatic sense. However, as convincingly argued by Pettigrove, third parties who declare forgiveness are not necessarily taking themselves to be victims. Strong evidence for this claim is that intimates connected to a victim may state their inability to forgive a perpetrator for what they did to *the victim* rather than to themselves. Genuine third-party forgiveness is distinguished, therefore, by having an other-regarding basis: the permissibility (or impermissibility) of such forgiveness is not grounded in the forgiver, but in an other (i.e., the victim). Paradigmatic forgiveness, by contrast, has a self-regarding basis: it is grounded within the individual who is forgiving (this difference will have explanatory importance detailed below). While it is beyond the scope of this discussion to independently establish the plausibility of third-party forgiveness, I aim to show that such forgiveness can be accommodated within the framework of negative partiality. I subsequently show how this analysis extends to self-forgiveness, providing a unified account of nonparadigmatic cases of forgiveness.

[26] For discussions, see Griswold, *Forgiveness*; Pettigrove, "The Standing to Forgive"; and Walker, "Third Parties and the Social Scaffolding of Forgiveness."

Several concerns must be kept in view when developing an account of third-party forgiveness. First, there is the risk of it undermining the significance of the *primary* victim's forgiving. In seeking forgiveness from the parents of a victim of assault, we might worry that the perpetrator has purchased inner peace in a way that inappropriately bypasses the moral imperative of the victim.[27] Perhaps reflecting this concern, the endorsement of third-party forgiveness typically comes with the caveat that the intimates of a victim should defer to the victim before offering up their own forgiveness. As a result, third-party forgiveness is distinguished by the fact that third parties typically do not *default* to having a prerogative. Second, third-party forgiveness risks allowing individuals far too disconnected from the initial wrong to be capable of forgiving. A successful account should provide some mechanism for limiting the scope of nonvictim forgivers.

To show how the framework of negative partiality accommodates third-party forgiveness, we must consider how negative partiality manifests in cases where our intimates have been wronged. For example, suppose that a small-town arsonist sets fire to a home owned by a local resident, Emily. Lucas, a local restaurateur and Emily's close friend, happens to have a project of working with and reintegrating convicts into the community by offering them employment. While the restaurateur's policy seems permissible, even admirable, we might take pause regarding the prospective employment of the arsonist. It is virtuous to reintegrate former convicts, but those connected to victims of crime should reasonably resist playing this role. Most importantly, Emily can plausibly complain about her friend offering such assistance. On this assessment, Lucas has an agent-relative basis for discounting the arsonist's well-being (i.e., a reason of negative partiality).

If third parties acquire duties of negative partiality when their intimates have been wronged, then they have an obligation to discount the interests of the perpetrator. As with conditional forgiveness, this obligation weakens merely to a prerogative when our intimates have either themselves forgiven the perpetrator and/or when relevant conditions have been met (repentance, apologies, etc.). For this reason, third parties do not *simply* have the prerogative to forgive: the standing to forgive arises when their duty weakens by meeting the relevant conditions. This picture addresses our concerns about third-party forgiveness. We might wonder who has the standing to forgive when they are not the immediate victim of a wrongdoing. The answer is to be found by looking at where we intuitively believe there are obligations, based on a relationship with the victim, to discount the interests of a perpetrator. This view rules out random strangers who merely "feel" a sense of association with the victim. Such individuals lack

[27] Dillon articulates a concern along these lines in "Self-Forgiveness and Self-Respect."

the standing to forgive insofar as they have no special duties of negative partiality. Second, this theory addresses the concern that third-party forgiveness will undermine the primary victim's forgiveness. Since victims and third parties have independent prerogatives of negative partiality, forgiveness by a third party does not undermine the primary victim's forgiveness. Third-party forgiveness, therefore, represents a second-best scenario when the primary victim is unwilling or cannot forgive. Such forgiveness reestablishes equality between the victim's relata and the perpetrator, even if it cannot establish equality among all relevant parties. Likewise, forgiveness by the primary victim does not *entail* forgiveness by one's intimates: it is still up to our relata to independently surrender their right to discount the interests of the perpetrator.

The explanation of why third-party forgiveness is distinct from paradigm forgiveness is that our intimates most plausibly have duties of negative partiality, rather than mere permissions. While this picture makes sense of third-party forgiveness, we could ask *why* duties of negative partiality arise in cases where our intimates have been victimized when our *own* victimization *typically* involves mere permissions. At this stage, I have offered several intuitive cases, but there is a deeper explanation for the pattern. It is to be found in the more primitive distinction between "self-regarding" reasons for action and "other-regarding" reasons for action. Common-sense morality recognizes a standing permission to discount our own interests *simply* because they are our own (self-sacrifice is *typically* meritorious and only invites moral criticism when at the expense of self-respect). For example, most people recognize a very weak duty to make *oneself* happy. By contrast, it is extremely uncommon to deny the *prima facie* obligation of beneficence.

The distinction between a self-regarding basis for action and an other-regarding basis helps to explain why third parties have duties of negative partiality, but primary victims (typically) do not. The *ground* of negative partiality when our intimates have been victimized is our intimate—i.e., the primary victim is the source of our reasons to be negatively partial to the perpetrator. Insofar as the ground of moral response is an other, negative partiality manifests as a duty; by contrast, our *own* victimization generally *permits* us to act with negative partiality, as the ground of this relation lies in ourselves (as with other cases of self-sacrifice, forgiveness will also often be supererogatory). When the primary victim offers *their* own forgiveness, they reestablish equality with the perpetrator; this act is one way of signaling to third parties that *their* forgiveness respects what is owed to the victim. However, the context of the initial forgiveness still matters: a primary victim who has forgiven but done so in a way that is inconsistent with self-respect does not provide a third party with an adequate basis for secondary forgiveness. Relatedly, a primary victim who forgives but otherwise signals a need for others to censure the perpetrator may thereby undermine the permissibility of third-party forgiveness. In sum, the permissibility of third-party forgiveness rests (at least in part) on whether the practice is consistent with appropriate consideration for the primary victim.

The self/other distinction also allows us to extend the account of forgiveness to the phenomenon of *self-forgiveness*. As with third-party forgiveness, we should be concerned by unreflective (or hasty) self-forgiveness. After all, would an agent not always *desire* to profit from self-forgiveness if possible? This concern risks stripping away the normative significance of this phenomenon.[28] The problem dissolves if self-forgiveness is instead seen as a special case of third-party forgiveness that applies to the perpetrator. Much like the victim, the perpetrator has special reasons for discounting their *own* interests in response to having victimized another. There is an affective analogue of this response—i.e., feelings of guilt—but, clearly, there are also implications for the actions of the perpetrator. This requirement is not best characterized as self-punishment but rather as a duty to avoid deriving benefit from the victim of their actions. Should, for example, our reforming arsonist seek networking advice from their *victim* as a means of furthering their reintegration into society? Plausibly not, even though the perpetrator can otherwise attempt to reintegrate. The possibility of self-forgiveness will, then, parallel cases of conditional forgiveness and third-party forgiveness—i.e., the perpetrator *begins* by having a duty to refrain from benefiting from their victims (i.e., a duty of negative partiality directed at themselves) and acquires the prerogative in light of relevant conditions being met (forgiveness by the victim *or* having sufficiently repented, apologized, etc.). The *duty,* in this case, is explained by the fact that it is grounded by agency of an other (i.e., the victim); after all, it is the *victim* who may complain when the perpetrator readily asserts self-forgiveness.[29] Moreover, as with third-party forgiveness, self-forgiveness never undermines the victim's forgiveness since neither affects the victim's prerogative.

[28] On Snow's view, self-forgiveness aims at "self-restoration." I endorse this idea in the sense that self-restoration can be understood as a *normative* phenomenon that allows one to have equal standing in the moral community. However, insofar as self-restoration is to be understood as a psychological/affective phenomenon, my view is distinct from Snow and others who understand self-forgiveness in these terms. For discussions, see Snow, "Self-Forgiveness"; Mills, "On Self-Forgiveness and Moral Self-Representation"; and Hughes, "On Forgiving Oneself."

[29] As Hughes argues, we can also forgive ourselves for wrongs to oneself ("On Forgiving Oneself"). Self-forgiveness in *this* sense would be analyzed in a way that approximates the paradigmatic case of forgiveness (i.e., insofar as one has only wronged oneself, there will be a prerogative to forgive.

This rough picture of self-forgiveness is continuous with third-party forgiveness yet stands in stark contrast to the dominant view, which understands self-forgiveness in descriptive terms (e.g., resolving negative psychic states, such as guilt). These views face similar challenges to those of descriptive accounts of paradigm forgiveness. A case of "wavering" about inner guilt could illustrate the point: have I forgiven myself if I experience a resurgence of guilt twenty years after the fact? Suppose instead that I *no longer* feel guilt but correctly believe that I *ought* to feel guilt. On a purely descriptive account, I *would* have forgiven myself so long as I have no such feelings (or other inner psychic trouble), but this seems intuitively untrue. Likewise, if I correctly believe that I *ought* to avoid deriving benefit from the victim of my action, I hardly count as having self-forgiven. Descriptive views cannot easily explain these observations.

6. Competing Performative Accounts

Understanding forgiveness as a performative akin to a promise is reflected by what Christopher Bennett terms the "alteration thesis," the idea that forgiveness *changes* a normative situation.[30] Bennett and Warmke have both recently argued that forgiveness waives obligations owed by the perpetrator to the victim, most notably the duty to compensate and apologize.[31] Bennett argues further that forgiveness may involve a recognition by the victim that the perpetrator has fulfilled their obligations, along with a commitment to treat the perpetrator in a corresponding manner (he terms this "redemptive forgiveness" since it redeems or recognizes redemption in the perpetrator). I clearly endorse the alteration thesis, understanding it in terms of surrendering the right to negative partiality. However, this difference in how the alteration thesis should be understood is significant. As argued below, I believe forgiveness does not alter the norms in the perpetrator (e.g., the duty to apologize and compensate) but should instead solely focus on the norms of victim.

7. Compensation, Apologies, and Promises

On Bennett's view, one function of forgiveness is to abrogate the duties to compensate and apologize to the victim (or cancel other secondary obligations

[30] Bennett, "The Alteration Thesis," 207.

[31] See Warmke, "The Normative Significance of Forgiveness." Hughes also suggests that forgiveness can be a performative, although he does not articulate a view about the moral change brought about by forgiveness ("On Forgiving Oneself"). Pettigrove also offers an early articulation of this view in "The Forgiveness We Speak."

acquired by the perpetrator in virtue of their wrongdoing).[32] This analysis raises the question of what forgiveness accomplishes when the perpetrator no longer owes anything to the victim of a wrongdoing. A problem arises, for example, when the perpetrator has *already* apologized or *already* offered compensation for their wrong. Since the *perpetrator* may take actions to execute these obligations, the perpetrator risks undermining the prerogative of forgiveness; there will be nothing left to forgive once the perpetrator's obligations are fulfilled. Bennett offers a novel solution to this problem by arguing that forgiveness alters the normative situation by an act of "redemptive" forgiveness, which plays the role of "acknowledging" that the perpetrator has fulfilled their obligations and generates an obligation in the victim to (going forward) treat the perpetrator *as if* they have fulfilled these obligations.

To assess this approach, I first consider whether forgiveness abrogates the duty to apologize. This understanding of the moral alteration brought about by forgiveness is somewhat striking when considering that providing an apology (or at least reiterating an apology) is often *prompted* by forgiveness. Such a reaction is difficult to interpret on the view that forgiveness waives the right to an apology. To illustrate, consider how two friends might navigate another debt that has been waived. Suppose April and Sheldon share lunch, and Sheldon tells April to "forget about it," thereby abrogating the duty of repayment. One fitting response to such an exchange is gratitude, but suppose instead that April attempts to repay Sheldon. In this case, the repayment is clearly an attempt to *reject* the abrogated duty—April does not *want* the debt cancelled, and repayment both acknowledges this fact and rejects the attempted abrogation. If forgiveness abrogates the duty to apologize, apologizing *post-forgiveness* suggests a "rejection" of the forgiveness, but clearly this is not the case; apologizing coheres with and reaffirms the rapprochement generated by forgiveness.

Another way in which forgiveness could alter the moral situation, according to Bennett and Warmke, is by waiving the right to compensation. To assess this claim, several ways of conceptualizing compensation should be distinguished. In the straightforward case, such as negligent damage to a vehicle, compensation has a price—i.e., the damages can be quantified in relatively uncontroversial financial terms. Compensation is harder to quantify when damages are abstract. The approach in a case of personal injury will typically involve placing a value on the loss of a bodily function, and while this compensation is said to make a person "whole," it is clearly metaphorical. Other abstract wrongs that give rise to the duty of compensation include "unjust enrichment" where

[32] See Twambly, "Mercy and Forgiveness," for another defense of the view that forgiveness involves waiving the right to compensation.

a perpetrator derives benefit from a person's property without permission or wrongs without any damages (e.g., harmless trespass). Notably, some serious affronts to a person are unlikely to be assessed primarily in terms of harm (e.g., the denial of the right to vote). In the aforementioned cases, a person is never literally made "whole" by compensation, and attaching a price to the transgression seems inherently contentious.

In the straightforward case where a person has suffered a loss with a *price*, I cannot see how forgiveness has any effect on the right to compensation. It seems perfectly consistent, for example, for a negligent driver to apologize and seek forgiveness, even if both parties recognize that the courts should assess and arbitrate an appropriate remedy for the accident. If forgiveness automatically gave up claims to compensation, forgiveness could only reasonably take place *after* a resolution of the case (or else these victims risk surrendering their claim). However, it is not extraordinary for the victims of such injuries to acknowledge forgiveness *and* seek restitution. Forgiveness may even be *predicated* on the expectation of restitution ("I know you're good for it"), implying a separation between the normative effects of forgiveness and requirements of restitution.

Claims of compensation can also be directed at wrongs with no corresponding price, such as the denial of political rights. Should forgiveness be understood to give up claims of reparations that result from these wrongs? This view seems at odds with the practice of reconciliation, which involves both forgiveness and forward-looking projects that attempt to redress wrongs. Consider, for example, the Truth and Reconciliation Report in Canada that simultaneously recognizes the right to reparations (and apology) for historical injustices and *seeks* forgiveness.[33] If the report ultimately led to what could be characterized as forgiveness, would the project of redress be abandoned? This conclusion is obviously against the spirit of the report. This idea goes back to much earlier discussions of reconciliation when Martin Luther King Jr. (MLK) detailed a path from forgiveness to love to reconciliation, all arguably in a manner that fits the alteration thesis:

> Forgiveness does not mean ignoring what has been done or putting a false label on an evil act. It means, rather, that the evil act no longer remains as a barrier to the relationship. Forgiveness is a catalyst creating the atmosphere necessary for a fresh start and a new beginning. It is the lifting of a burden or the canceling of a debt.[34]

[33] See Truth and Reconciliation Commission of Canada, "Honouring the Truth, Reconciling for the Future."

[34] King, *A Gift of Love*, 47.

The "cancellation" of a debt is clearly a notion friendly to the concept of forgiveness as a normative power akin to promising, but MLK never characterized this debt in terms of compensatory justice:

> When white Americans tell the negro to lift himself by his own bootstraps they don't look over the legacy of slavery and segregation. I believe we ought to do all we can and seek to lift ourselves by our own bootstraps but it's a cruel jest to say to a bootless man that he ought to lift himself by his own bootstraps. And many negroes by the thousands and millions have been left bootless as a result of all of these years of oppression and as a result of a society that has deliberately made his color a stigma and something worthless and degrading.[35]

MLK clearly called for forgiveness as a way of repairing a relationship shaped by historical wrong, but simultaneously pressed for claims of compensation. These concurrent claims should strike us as perfectly consistent, but they are incompatible with the claim that forgiveness gives up all claims that arise in virtue of a wrong.

The focus on compensation and apologies in competing articulations of the alteration thesis also raises difficulties for the interpretation of nonparadigm cases of forgiveness. First, I know of no attempt to advance the idea that an individual who wrongs another (or themselves) has a duty to apologize or compensate themselves. It will therefore be difficult to accommodate self-forgiveness within this framework. Second, while apologies might be owed to secondary victims in *extreme* cases of wrongdoing, it seems unlikely to arise in cases of moderate wrong (e.g., the plagiarism case or infidelity). In these cases, it is likewise difficult to see how Warmke or Bennett will capture third-party forgiveness.

8. Redemptive Forgiveness

Apart from waiving the right to compensation or an apology, Bennett offers the unique suggestion that forgiveness can take the form of "redemption," which involves *recognizing* that the perpetrator has fulfilled their duty to apologize, compensate, etc., and committing to treat them as if these obligations have been fulfilled. This commitment is a "change of stance … thought of as "bracketing" at least some of the normative effects of that particular wrongdoing as a basis for one's relationship with the wrongdoer and making it the case that one will wrong him should one go back on one's undertaking and start to treat him as

[35] King, interview by Sander Vanocur.

one who stands under those obligations of which he is now free."[36] Redemptive forgiveness is susceptible to a range of problems that emerge when considering how we ought to respond to special obligations that have been fulfilled. Consider, again, the case of paying back a loan. April owes Sheldon twenty-five dollars for lunch, and April repays the loan in a timely fashion. Once April's debt has been executed, it seems strange to say that Sheldon is in a position of choosing whether to grant an "acknowledgment" that the debt has been repaid. Sheldon need not *declare* "April's debt has been repaid," but Sheldon certainly cannot *deny* the repayment (if anybody asks), and Sheldon cannot do activities typically associated with being owed a debt. It would, for example, be impermissible for Sheldon to *demand* repayment. In broad terms, once a debt has been fulfilled, the former obligee must act as if the debt is fulfilled. Redemptive forgiveness, therefore, seems unable to make a normative difference of the kind needed: it cannot reestablish moral equality. Once the debt has been repaid, the parties are equal.

It is true that a further *commitment* to treat the perpetrator in the appropriate fashion changes the moral situation by introducing a *stronger* obligation to treat them with respect, but such a commitment is not a matter of reestablishing moral equality. We generally stand in a relation of moral equality *regardless* of our commitment to doing so. Insofar as we need an explanation of how forgiveness reestablishes moral equality and "raises" the standing of perpetrator, a commitment to respect them seems insufficient. This view stands in contrast to the position that forgiveness involves surrendering rights to negative partiality, which provides a concrete interpretation of how forgiveness elevates the moral standing of the perpetrator.

9. Descriptive Accounts Revisited

Through an initial critique of descriptive accounts of forgiveness, I motivated the idea that forgiveness is normatively significant. With the positive view now tabled, it is worth revisiting how these different approaches may be related.[37] Must we view normative and descriptive accounts as mutually exclusive, and how (given a normative understanding of forgiveness) ought we to interpret the progress that has otherwise been made on the psychological and behavioral dimensions of forgiveness? Despite my claim that a purely descriptive account of forgiveness leaves out elements of this phenomenon, there is more harmony (or, at least, *potential* harmony) between descriptive and normative accounts

[36] Bennett, "The Alteration Thesis," 219.

[37] I am grateful to several anonymous reviewers who raised questions/objections regarding the relationship between descriptive and normative theories explored below.

of forgiveness than may initially appear. In this brief section, I detail several ways that these views could be connected; given the complexity of this issue, I remain agnostic as to the connection we *ought* to embrace.

My central concern with descriptive accounts of forgiveness is that behavioral and psychological changes are insufficient to capture *some* practices surrounding forgiveness. This narrow claim does not eliminate the potential for psychological and behavioral changes (as discussed in the extant literature) to play a role in successful acts of forgiveness. One approach that connects these normative and descriptive views is deflationary and merely takes the descriptive changes in a subject to play a causal role in bringing about normative changes (these latter changes being identified as forgiveness proper). For example, overcoming resentment for a moral reason or coming to see the perpetrator in a better light may *motivate* the victim to surrender rights held against the perpetrator. Insofar as such psychological changes are neither necessary nor sufficient for altering the norms of the relationship with the perpetrator, this proposal would significantly diminish the significance of descriptive views.

A stronger and perhaps more plausible account takes there to be an intrinsic connection between psychological changes and the normative effects of forgiveness. To illustrate by analogy, consider the idea that promises surrender the right to refrain from acting in ways that are inconsistent with the content of the promise. While this normative effect may be central to promises, the conditions of a successful promise plausibly include descriptive conditions for the alteration to succeed. For example, it may be a requirement of a promise that the promisee hears, understands, and acknowledges the promise. It may likewise be the case that in order for forgiveness to succeed—i.e., a successful surrendering of the right to negative partiality—the victim must undergo certain psychological changes (some of which may be in line with what has been examined in the literature). It seems implausible that a victim can successfully surrender a right to negative partiality if they have forgotten the wrong; it is much more plausible that a victim can surrender rights through a process that involves a reevaluation of the perpetrator's character. If this reevaluation is *required* for forgiveness, then there will be an intrinsic link between descriptive and normative accounts of forgiveness. On this view, both descriptive and normative conditions may end up being necessary for forgiveness.

There are further ways of preserving descriptive and normative accounts of forgiveness through conceptual pluralism. If we conclude that both accounts provide necessary and sufficient conditions of forgiveness, we can retain consistency only by expanding the conceptual sphere and admitting that there is more than one sense in which a person can forgive. Such a move comes at the cost of parsimony but may ultimately be the most accurate way of dividing up the class of activities that can legitimately be called "forgiveness." A more parsimonious way of capturing the pluralistic sentiment might draw a distinction between a minimum threshold of forgiveness being met and the *ideals* of forgiveness. While an essential element of forgiveness could include surrendering the right to resent, the *actual* overcoming of resentment could be taken to represent an ideal of forgiveness. Many psychological/behavioral changes fit a similar bill, such as the resumption of normal relations with the offender or having goodwill toward the offender. These changes might be classified as the ideals of forgiveness rather than necessary elements of forgiveness.

Related to the issue of mutual exclusivity, it may be asked why my position cannot simply be reimagined as a new *descriptive* theory of forgiveness. After all, the view I have articulated may seem closely related to a candidate for a description of the psychosocial processes that, in fact, unfold when a person forgives—i.e., the victim at one point assigned less weight to the interests of the perpetrator and subsequently ceased to do so. Why could these factual changes not be understood to capture forgiveness, and if so, what is the appeal of adopting Bennett's "alteration thesis"? To understand my concern with this position, consider a victim who declares their forgiveness. If the victim has altered their attitudes toward the perpetrator, the statement will reflect a genuine change that occurred, and if they have failed to do so, the statement will be either mistaken or dishonest. Now, suppose that going forward, the victim *continues* to discount the interests of the perpetrator. On a descriptive view, the victim has failed to accurately report their attitudes, but apart from this inaccurate (or dishonest) reporting, they have done nothing wrong. The problem with this position is that it fails to capture the sense in which the perpetrator can legitimately expect the victim to act differently. By *declaring* forgiveness and acting otherwise, the victim did not merely fail to report their attitudes, they failed to live up to an obligation that was incurred through their declaration. This is the sense in which forgiveness has a performative dimension that alters moral norms, one akin to how promises bind through declarations.

Yet, something may seem amiss in the above example: how can we say that a person has forgiven another individual if they continue to discount their interests? It may seem strange to say without hesitation that an individual who has performed an act of *revenge* against another can count as having forgiven that same person. Here, I believe two aspects of forgiveness are in tension. On the one hand, we tend to hold someone who forgives another *accountable* for the fact that they have forgiven and criticize them if they fail to act in accordance with a declaration of forgiveness. On the other hand, we may be reluctant to describe someone who fails to act in accordance with the norms of forgiveness as having *truly* forgiven. These uses are inconsistent. If forgiveness has genuinely not occurred, there should be nothing to criticize about the person who acted inconsistently with the norms of forgiveness. Which of these two uses should prevail? To shed light on this issue, it may be worth comparing another practice that involves a similar duality in the use of a concept. Suppose that Justin's best friend Jess has failed to live up to a norm of friendship (e.g., suppose that Jess desperately needs a ride to a job interview and Justin refuses because he *never* skips leg day at the gym). Jess might very well assert that "she thought Justin was her friend," implying that he was not her friend. However, if Justin and Jess have what would otherwise be described as a long and intimate relationship, it would be more plausible to say that his act is impermissible *because* of their friendship. After all, without recognizing the existence of the friendship, it would be difficult to explain *why* anything problematic occurred (the phrase "you're no son of mine" likewise gives rise to this duality: the statement presupposes the relationship it seeks to undermine). When a person declares forgiveness and acts inconsistently with the declaration, we might very well say that they have not *truly* forgiven. I would read this case in one of two ways. We are either saying that they have failed to live up to the *norms* of forgiveness (much like the case of friendship), *or* we are recognizing that a further felicity condition of forgiveness (as described in the previous section) has not been met. On either reading, my position cannot be transformed into a purely descriptive view.

10. Conclusion

I have argued in the spirit of Bennett and Warmke that forgiveness brings about a moral alteration akin to a promise. In contrast with previously established views, the scope of the alteration brought about by forgiveness should focus on the class of actions that may be performed by the victim. The attraction of this view lies in its ability to capture the core elements of forgiveness, such as its personal nature, its distinction from excusing or justification, its normative significance, and its fit with varying types of forgiveness (conditional, unconditional, self, and third-party). This broad explanatory power derives from the simple proposition that negative partiality represents a relationship of moral distance, and forgiveness acts to eliminate this distance.[38]

BC Office of the Ombudsperson
joshua.brandt@utoronto.ca

[38] This paper was written while I was working as a faculty member at the University of Toronto. Its central ideas were formulated while I was writing my thesis under the supervision of Thomas Hurka. I am grateful for Tom's help in working through these ideas as well as the thoughtful feedback and encouragement provided by Amy Mullin, Andrew Sepielli,

References

Allais, Lucy. "Wiping the Slate Clean: The Heart of Forgiveness." *Philosophy and Public Affairs* 36, no. 1 (Winter 2008): 33–68.

Bennett, Christopher. "The Alteration Thesis: Forgiveness as a Normative Power." *Philosophy and Public Affairs* 46, no. 2 (Spring 2018): 207–33.

Brandt, Josh. "Negative Partiality." *Journal of Moral Philosophy* 17, no. 1 (February 2020): 33–55.

Broad, C. D. "Self and Others." In *Broad's Critical Essays in Moral Philosophy,* edited by David Cheney, 262–82. New York: Humanities Press, 1971.

Cottingham, John. "Partiality, Favouritism and Morality." *Philosophical Quarterly* 36, no. 144 (July 1986): 357–73.

Davis, Jeremy. "Scope Restrictions, National Partiality, and War." *Journal of Ethics and Social Philosophy* 20, no. 2 (August 2021): 144–67.

Dillon, Robin S. "Self-Forgiveness and Self-Respect." *Ethics* 112, no. 1 (October 2001): 53–83.

Dworkin, Ronald. *Law's Empire*. Cambridge, MA: Harvard University Press, 1986.

Griswold, Charles. *Forgiveness: A Philosophical Exploration*. New York: Cambridge University Press, 2007.

Haber, Joram Graf. *Forgiveness*. Lanham, MD: Rowman and Littlefield, 1991.

Hieronymi, Pamela. "Articulating an Uncompromising Forgiveness." *Philosophy and Phenomenological Research* 62, no. 3 (May 2001): 529–55.

Hughes, Paul M. "On Forgiving Oneself: A Reply to Snow." *Journal of Value Inquiry* 28, no. 4 (December 1994): 557–60.

Hurka, Thomas, and Esther Shubert. "Permissions to Do Less Than the Best: A Moving Band." In *Oxford Studies in Normative Ethics,* vol. 2, edited by Mark Timmons, 1–27. Oxford: Oxford University Press, 2012.

King, Martin Luther, Jr. *A Gift of Love: Sermons from* Strength to Love *and Other Preachings*. Boston: Beacon Press, 2012.

———. Interview by Sander Vanocur. *After Civil Rights: Black Power*. NBC, June 11, 1967. https://www.youtube.com/watch?v=2xsbt3a7K-8.

Kolnai, Aurel. "Forgiveness." *Proceedings of the Aristotelian Society* 74, no. 1 (June 1973): 91–106.

and Sergio Tenenbaum, the other members of my thesis committee. Earlier versions of this paper were presented at the Canadian Philosophical Association's Annual Meeting, the Oxford Partiality Workshop, and the Department of Philosophy at the University of Toronto. I am grateful for commentaries given by Kathryn Norlock and Aart Van Gils. I would also like to thank Manish Oza and Reuven Brandt for their helpful comments on earlier drafts.

Lange, Benjamin. "Other-Sacrificing Options." *Philosophy and Phenomenological Research* 101, no. 3 (November 2020): 612–29.

Milam, Per-Erik. "Reasons to Forgive." *Analysis* 79, no. 2 (April 2019): 242–51.

Mills, Jon K. "On Self-Forgiveness and Moral Self-Representation." *Journal of Value Inquiry* 29, no. 3 (September 1995): 405–6.

Murphy, Jeffrie G. *Getting Even: Forgiveness and Its Limits*. New York: Oxford University Press, 2003.

Murphy, Jeffrie, and Jean Hampton. *Forgiveness and Mercy*. Cambridge: Cambridge University Press, 1988.

Newberry, Paul A. "Joseph Butler on Forgiveness: A Presupposed Theory of Emotion." *Journal of the History of Ideas* 62, no. 2 (April 2001): 233–44.

Pettigrove, Glen. "The Forgiveness We Speak: The Illocutionary Force of Forgiving." *Southern Journal of Philosophy* 42, no. 3 (Fall 2004): 371–92.

———. "The Standing to Forgive." *Monist* 92, no. 4 (October 2009): 583–603.

Scheffler, Samuel. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford: Clarendon Press, 1982.

Snow, Nancy E. "Self-Forgiveness." *Journal of Value Inquiry* 27, no. 1 (January 1993): 75–80.

Stroud, Sarah. "Permissible Partiality, Projects, and Plural Agency." In *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*, edited by Brian Feltham and John Cottingham, 131–49. Oxford: Oxford University Press, 2010.

Truth and Reconciliation Commission of Canada. "Honouring the Truth, Reconciling for the Future: Summary of the Final Report of the Truth and Reconciliation Commission of Canada." 2015. https://publications.gc.ca/site/eng/9.800288/publication.html.

Twambley, Paul. "Mercy and Forgiveness." *Analysis* 36, no. 2 (January 1976): 84–90.

Walker, Margaret Urban. "Third Parties and the Social Scaffolding of Forgiveness." *Journal of Religious Ethics* 41, no. 3 (September 2013): 495–512.

Warmke, Brandon. "The Normative Significance of Forgiveness." *Australasian Journal of Philosophy* 94, no. 4 (2016): 687–703.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 1 · MARCH 2024

https://doi.org/10.26556/jesp.v27i1.2614

© 2024 Authors

THE PROCEDURE OF MORALITY

Ori J. Herstein and Ofer Malcai

DOES morality have a procedure? In some normative realms, such as law, procedural norms are commonplace. In fact, given that law inherently involves lawmaking and law-applying institutions, law's vast web of procedural norms seems almost inevitable. But what about normative realms that are not inherently institutional? Namely, are procedural norms part of moral discourse?

We argue that procedural norms akin to those found in the law are commonplace in morality as well, so much so that you could say that, like law, morality too has a "procedural branch"—what David Enoch has referred to as the "underexplored territory of the procedural law of morality."[1]

The view that morality has a procedure is not only underexplored but controversial.[2] In fact, the mere term "moral procedure" can sound almost oxymoronic. After all, morality lacks institutions and may seem—in its very essence—"substantive" all the way down. Indeed, some are skeptical about procedural moral norms or related notions such as procedural moral rights and duties. For example, Christopher Wellman has rejected the existence of

[1] Enoch, "In Defense of Procedural Rights (or Anyway, Procedural Duties)," 49.

[2] To be sure, there are discussions in moral philosophy in the neighborhood of our inquiry. For instance, the distinction between "substance" and "procedure" seems related to the distinction between "form" and "content," such as in debates over Immanuel Kant's formalist categorical imperative (e.g., Rawls, *Lectures on the History of Moral Philosophy*, 164–70; and O'Neill, *Acting on Principle*, 111, 136–93). However, our characterization of procedural moral norms is independent of any particular position about ethics or political morality (e.g., Kantian and Rawlsian positions). Accordingly, we do not offer a specific procedural mechanism for generating basic normative principles or practical prescriptions, such as John Rawls's "original position" (Rawls, *A Theory of Justice*) or Kant's categorical imperatives, but rather provide general conditions for what it is for a norm to be "procedural."

The literature on procedural justice is even more directly relevant to our inquiry (e.g., Solum, "Procedural Justice"), as it identifies some characteristics of norms that are pretheoretically considered as "procedural." However, much of the discussion of procedural justice in moral philosophy focuses on the normative properties of certain procedures (e.g., on what makes them just or fair) rather than on the procedural properties of norms, which is our basic concern here; hence the title of the paper: "The Procedure of Morality."

procedural moral rights, arguing that procedural rights, such as the right against punishing a person without first establishing her guilt via due process, do not exist pre-institutionally.[3] Yet others do not reject the very existence of procedural rights but obscure the distinction between procedure and substance, arguing that "procedural rights just are substantive rights."[4]

In this paper, we offer a theory of procedure for normative domains. We begin by defining what it is for a norm to be "procedural," suggesting that procedural norms are a distinct normative kind with identifiable general characteristics, distinguishing them from the category of "substantive norms." The paper is largely conceptual rather than normative, offering insights into the structure and internal tensions of moral discourse. Methodologically, we first test our theory of procedure on instances of legal procedure, ensuring that our theory adequately captures what are commonly and pretheoretically considered paradigmatic instances of the procedural branch of the law. Then, moving from law into the domain of morality, we identify an incompatibility between procedural and substantive norms, raising the worry that procedural *moral* norms are conceptually paradoxical or, at the very least, morally untenable. We then tackle these objections, vindicating the view that morality has a procedure.

The paper is structured as follows. Section 1 is devoted to our account of what makes a norm procedural. Section 2 demonstrates how our account captures norms of *legal* procedure and, along the way, identifies three central types of procedural norms. Section 3 articulates three objections to the notion that morality has a procedure—the *no-institution objection*, the *conceptual objection*, and the *moral objection*. That section then addresses the first two of these objections, leaving the third objection to section 4, where we respond through counterexamples of familiar and intuitive moral norms exhibiting the features of procedural norms according to the account presented in section 1.

1. Characterizing Procedural Norms

We conceptualize procedural norms as involving three related features.[5] Namely, they are *second-order* norms, they are about *how to engage* with other norms, and they are *outcome neutral*.

[3] Wellman, "Procedural Rights."

[4] Alexander, "Are Procedural Rights Derivative Substantive Rights?"

[5] We do not define the term "norm." Rather, we use "norm" more loosely, broadly referring to propositions incorporating normative predicates or operators, such as "ought," "permissible," or "justified."

1.1. Second Orderness

Roughly speaking, second-order norms are norms about other norms, that is, providing reasons related to other norms (or to the reasons provided by those other norms). Generally, procedural norms set a normative framework for dealing with other (typically substantive) norms.[6] For example, rules of evidence are second-order norms in that they instruct courts on how to decide whether the relevant substantive norms governing the case (e.g., of criminal law, torts, etc.) have been violated.[7]

Like first-order norms, second-order norms provide agents with reasons for action (or emotion, or belief), except that unlike first-order norms, which determine the normative status of nonnormative facts (such as actions or states of affairs), we hold that when it comes to second-order norms, the thing whose normative status is at stake is itself characterized in normative terms.[8] For example, "it is morally wrong to punish a person for an action that is morally permissible" is a second-order norm because the thing whose normative status is at stake—namely, the act of punishing a person for an action that is morally permissible—is characterized in normative terms ("morally permissible"). More formally, second-order norms are expressible by sentences that include a *normative term* within the scope of a normative predicate or operator.[9] For example, in the aforementioned norm of punishment, the term "morally permissible" is embedded within the scope of the predicate "morally wrong."

Our definition of "second-order norm" is stipulative, and it does not depend on any correspondence to other uses of the term "second-order."[10] Nevertheless, we believe that our characterization of second-order norms captures an

[6] For similar characterizations of procedural norms, see Malcai and Levine-Schnur, "Which Came First, the Procedure or the Substance?" 69; see also Rosenthal, "What Decision Theory Can't Tell Us about Moral Uncertainty," 3089–90.

[7] For instance, rule 403 of the Federal Rules of Evidence (as amended December 1, 2022) states:

> The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence.

This is a second-order norm given that it is about how courts ought to decide on the parties' compliance with the law's substantive norms.

[8] Malcai, "Second-Order Propositions and Metaethical Neutrality."

[9] Examples of normative terms include "good," "bad," "right," "wrong," and "permissible."

[10] In particular, what we call "second-order norms" do not require second-order logic for their formulation. For example, second-order norms can perhaps be expressed by conditional sentences in first-order logic.

important feature of the normative discourse (be it moral, epistemological, or legal).

1.2. About the "How"

Procedural norms are one kind of second-order norm. What colors a second-order norm as procedural is the providing of reasons bearing on the "how" of agents' *engagement* with other norms (or with the reasons provided by other norms). By "engagement" we have in mind something very general, including any instance of agency responding to norms, deliberating on norms, applying norms, and forming norms. These second-order norms warrant the label "procedural" because they are, broadly speaking, about the process of one's engaging with other norms.

1.3. Outcome Neutrality

Relatedly, procedural norms are in a sense outcome neutral. As norms about how to engage with other norms, procedural norms are about the process of such engagement as opposed to the normative outcome of the engagement itself.

Now, clearly, in bearing on the process of engagement with other norms, procedural norms can also impact the normative outcome of such engagement. Yet, what is crucial to notice is that they only do so *indirectly*, as the procedural norm itself does not bear on the matter. For example, in determining whether it ought to convict a defendant, a court ought to deploy the "beyond reasonable doubt" standard of persuasion. Now, this evidentiary norm can of course impact whether or not a defendant ought to be convicted, but being a second-order norm, it does not bear on which verdict the court ought to reach; it only instructs the court on how to engage with those norms of (substantive) criminal law that *do* determine the normative status of the defendant's actions.

To fully appreciate the outcome neutrality of procedural norms, consider the *non*procedural second-order norm that we encountered at the outset: "punishing a person for an action that is morally permissible is morally wrong." While this norm is a second-order norm—it relies on another norm to set its scope (namely, on those norms that determine the moral permissibility of actions)—it is *not* a procedural norm, as it does not bear on the process of one's engagement with any other norm. Rather, it directly determines the moral status of the outcome, namely, the appropriateness of the punishment, regardless of the appropriateness of the process by which this outcome is arrived at. As such, this norm is *not* outcome neutral. Accordingly, it is a "substantive (second-order) norm" and not what we here call a "procedural norm."

One could presumably object to the existence of procedural norms, since there is something contradictory in a norm that is agnostic about what it counts in favor of. Namely, given that norms by definition provide reason for ϕ-ing, norms are by their very nature not neutral as to ϕ-ing. The outcome neutrality of procedural norms might be taken to conflict with this truism.

Happily, this worry is easily dealt with, as it involves a mischaracterization of the outcome neutrality of procedural norms. Outcome neutrality does not make procedural norms normatively inert. Procedural norms are not agnostic vis-à-vis what they *do* bear on directly, which is a certain *form* of engagement with another norm. Accordingly, the fact that a procedural norm is neutral on the normative outcome of the form of engagement that it counts in favor of does not entail that *that* procedural norm is normatively agnostic through and through. Thus, the outcome neutrality of procedural norms does not negate their normative nature.

1.4. The Procedure of Morality, not the Morality of Procedure

To avoid confusion, before turning to elaborate on a handful of different types of procedural norms, it is worth distinguishing our concept of "procedural norm" from other possible meanings of the term. In natural language, the term "procedural norm" comfortably encapsulates types of norms exceeding the philosophical type that we have in mind here. In fact, there are numerous norms advising or prescribing procedures for performing certain actions in a certain manner or order. Examples vary from surgical protocols to the sequenced routines that parents instill in their young children for going to sleep (e.g., bathing, donning pajamas, brushing teeth, then enjoying a lullaby).

In some sense, such norms are also "procedural." Beyond the fact that they advise or prescribe a certain procedure, such norms also exhibit certain procedural-like features. For one, these norms are about the process and the "how" of reaching certain ends or performing certain overarching actions other than the actions the norms themselves directly prescribe. For example, a parental directive to first bathe, then dress, then brush, and so on prescribes a sequence of actions comprising a process for how children are to perform the overarching action of turning in for the night.

Yet notwithstanding their procedural-like features, these procedure-prescribing norms differ significantly from those norms that we label "procedural." Notice first that procedure-prescribing norms such as the parental bedtime directive are first-order norms. While referring to such norms as "procedural" might be compatible with natural language, these norms, as far as we can see, do not raise unique philosophical questions similar to those raised by the norms that we label "procedural." Thus, incorporating these types of procedural-like norms into our picture of procedural norms risks drowning out the normative phenomenon that we aim to elucidate.

We therefore propose a distinction between what we call "procedural norms," which are the type of norms that we are interested in here (namely, norms about how to engage with other norms), and first-order norms that prescribe following a certain procedure. Although somewhat overlooked in moral philosophy, the fact that this distinction captures a unique normative kind is sharply reflected in the prevailing *legal* distinction between substance and procedure. For example, a legal norm requiring physicians to follow a certain protocol when disclosing medical information to a patient is clearly part not of the procedural branch of the law but rather of the relevant substantive law determining physicians' duties and patients' rights. In contrast, what we call "procedural norms" are not merely norms prescribing conduct plausibly labeled "procedural"; rather, they *are* procedural, as they embody the procedure for how to rightly engage with other norms. For example, courts ought to rely on expert testimony about the medical practice and state of the art in determining what is required by the legal standard of care in cases of medical malpractice; this legal norm is procedural because it prescribes *how* to go about determining what the standard of care under (substantive) negligence law is.

A possible objection to our position is to argue that all instances of what we call "procedural moral norms" can also be formalized as first-order norms of only one normative predicate or operator. For example, one might reformulate "in determining whether it *ought to* convict a defendant, a court ought to deploy the 'beyond reasonable doubt' standard of persuasion," as "in determining whether *to* convict a defendant, a court ought to deploy the 'beyond reasonable doubt' standard of persuasion." If so, our formulation of procedural moral norms is artificial, as such norms are reducible to straightforward first-order moral norms prescribing procedures.

Nevertheless, this reformulation in first-order terms obscures the normative quality of the practical matter at stake. Arguably, when a normative system (such as law) prescribes ϕ-ing, it actually prescribes that one *ought to* ϕ (according to that system). For instance, when deciding to convict an accused, the judge is following norms prescribing that under the circumstances the accused *ought to* be convicted according to the law. In contrast, were the judge following the demands of a violent mob to convict the accused, she then might indeed act on a reason for how to rule that is not laden with a norm about how she ought to rule (according to the law).[11]

[11] To concretize, we return to this last objection when discussing one of our examples (section 4.3).

2. TYPES OF PROCEDURAL NORMS

As norms about *how* to engage with other norms, procedural norms vary in the type of engagement that they prescribe. Below, we detail a few central examples, grouping them into three rough categories of norms exhibiting the three procedural characteristics detailed above. We deliberately draw these initial examples from the law, in which the existence of procedural norms is widely recognized.

2.1. Norms of Deliberation

Some procedural norms directly guide one's *deliberation* on other norms, thus bearing on the process of *reasoning* and on the *decision-making* itself. An example is the aforementioned judicial standard of persuasion. For instance, battery—be it the crime or the tort—mostly comprises similar elements.[12] This similarity notwithstanding, criminal law and tort law differ significantly in their procedures. In particular, the standard for persuading courts of defendants' civil liability ("preponderance of the evidence") is lower than the standard for persuading criminal courts of defendants' guilt ("beyond reasonable doubt"). Thus, criminal law and tort law differ less in their similar substantive norms of battery and more in their procedural frameworks governing the court's decision-making processes when applying those largely similar (substantive) norms.

2.2. Norms of the Application of Norms

Procedural norms can also bear on the *manner and means of applying* another norm. Such norms govern *practical aspects* of the process of engaging with other norms. Criminal procedure is chock-full of examples, such as criminal defendants' Sixth Amendment rights bearing on the form and management of criminal trials:

> In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed . . . to be confronted with the witnesses against him; to have compulsory process for obtaining witnesses in his favor, and to have the Assistance of Counsel for his defence.[13]

[12] E.g., in California, criminal battery is any willful, unlawful use of force or violence on the person of another (People v. Pennington, 3 Cal. 5th 786, 792 (Cal. 2017)), while the California tort of battery comprises intentional, unlawful, and harmful or offensive contact by one person with the person of the other (Barbara A. v. John G. 145 Cal. App. 3d 369 (1983)).

[13] U.S. Constitution, art. 1, sec. 5.

These rights are procedural: they are second-order norms, as they are about other norms (of substantive criminal law); they are about the practicalities (a speedy trial by an impartial jury, etc.) of how to apply those other norms; and finally, they are outcome neutral, bearing on the form of the criminal trial and not (at least not directly) on its normative outcomes.

2.3. Forming, Shaping, and Validating Norms

Another mode of procedural norms involves rules regulating the forming and shaping—that is, creating, altering, or annulling—of other norms. In law, these are akin to H. L. A. Hart's "rules of change."[14] Consider, for example, the United States Congress's rules for passing legislation that, among other conditions, require that a bill pass by a simple majority in the House of Representatives.[15] This legal norm is procedural: it is second-order, given that it is about (the creation of) other legal norms; it controls *how* Congress legally ought to form new legal norms; and it is outcome neutral, given its agnosticism about the content of those new norms.

3. IS MORALITY NOT SUBSTANTIVE ALL THE WAY DOWN?

Now that we have a handle on what procedural norms are and are not, we turn to tackle three objections to the view that *morality* exhibits such norms.

3.1. There Are No Moral Institutions

Our discussion of procedural norms has thus far gravitated toward the law. This is not surprising. Modern legal systems invariably involve institutions, and institutions generally incorporate procedures as part of their operations and even their constitution. Moreover, typically such procedures are governed by norms, certainly in the case of complex social institutions. Finally, arguably the core function of legal institutions is the formation and application of legal norms. Thus, characteristically procedural norms govern the law-forming and law-applying functions of legal institutions. Accordingly, legal systems typically involve two kinds of norms: substantive norms, which are usually directed outwardly toward the citizenry, and procedural norms, which govern how legal institutions (and those involved with them) ought to engage with (e.g., apply, deliberate, form, shape, or validate) the law's substantive norms.

[14] Indeed, we view Hart's project partially as adumbrating law's foundational procedural norms (*The Concept of Law*, 115–17).

[15] U.S. Constitution, amend. XI.

Morality is crucially different from the law in this respect. Simply put, there are no institutions of morality, at least none equivalent to those found in the law, such as "moral legislators," "moral courts," and "moral advocates." Accordingly, given the apparent tie between procedural norms and institutions, and given that morality is institution-free (in the relevant sense), there is at least some reason to doubt our notion of procedural *moral* norms.

One possible response to this objection is to reject—on metaethical grounds—the disanalogy between law and morality regarding the role of institutions and procedures. For example, what some label "metaethical constructivism" holds that the criteria for the truth value of moral propositions are dependent on a certain (actual or hypothetical) procedure, such as, for example, what rational agents would agree to under some set of specified conditions.[16] We will not pursue this line of response. Our conception of procedural moral norms is agnostic about the metaethical debate over the role of procedure in determining the truth value of moral propositions. Our concern here is whether there are identifiable procedural norms *within* moral discourse, regardless of the metaethical question of what the criteria for what falls within that discourse are. The arguments proposed here for the existence of procedural moral norms are thus compatible with both constructivist and nonconstructivist metaethical views.

Our response to the institutional objection is, first, that, conceptually, there is nothing institutional in our tripod account of procedural norms as outcome-neutral second-order norms about how to engage with other norms; and, as argued below, this tripod account of procedural norms is conceptually sound (section 3.2.1). Second, transitioning from the conceptual response to the moral, while some of our examples of procedural moral norms are justified only assuming some institutional backdrop, the justifications of others (e.g., epistemic moral norms) are institution-free (section 4).

3.2. The Conceptual and Moral Objections to Procedural Moral Norms

There are, however, at least two deeper objections to the view that morality involves a procedure, which we label the *conceptual objection* (section 3.2.1) and the *moral objection* (section 3.2.2). These objections arise out of the outcome neutrality of procedural norms. Nevertheless, while acknowledging the complex and even somewhat paradoxical nature of outcome-neutral procedural *moral* norms, we argue that these complications do not rule out procedural norms from populating morality. As a conceptual matter, it is the second

[16] For this characterization of constructivism, see Enoch, "Can There Be a Global, Interesting, Coherent Constructivism about Practical Reason?"; Bagnoli, "Constructivism in Metaethics."

orderness of procedural norms that unlocks the conceptual possibility of such norms. And, as a moral matter, while we recognize the possible clash between procedural and substantive moral norms, we do not think that such a clash wholly undermines the former. Indeed, this friction between the procedural and the substantive branches of morality is not a "bug" in our theory but rather a vital feature of morality. Or so we will argue.

3.2.1. The Conceptual Objection

Recall, procedural norms are outcome neutral; they bear on the form or process of engaging with norms, not on the normative outcome of such engagement. One may object that such outcome neutrality is *suapte natura* inconsistent with *moral* norms. Arguably, a moral norm counting in favor of ϕ-ing does so in virtue of the morally relevant (factual) properties of ϕ-ing. For example, a norm counting in favor of praising one's student does so by virtue of whether the student did anything praiseworthy. This supervenience of the moral status of ϕ-ing on the morally relevant properties, which can seem inherent to moral norms, might prove incompatible with the existence of outcome-neutral procedural norms. That is because the outcome neutrality of procedural norms entails myopia toward what appear to be the morally relevant reasons counting for or against ϕ-ing. Accordingly, following a procedural norm can seemingly result in a moral prescription to do something that is itself morally wrong—which has the air of paradox.

Suppose there is a certain procedural norm requiring that you apply a certain decision-making procedure P in order to decide whether, as a moral matter, you ought to ϕ. If ϕ-ing is morally required, it seems to be so in virtue of some morally relevant properties of ϕ-ing *itself*—it increases others' well-being, it promotes equality, it folds out of a good will, and so on. This is so regardless of whether or not you decided that you ought to ϕ by applying procedure P. Conversely, if independently of applying P, ϕ-ing itself is morally wrong—for example, it is unjust or it causes suffering—it seems that it must remain wrong even if you decided that you ought to ϕ by applying procedure P. Indeed, following a procedural norm can result in an apparent paradox: a moral prescription to do something that is itself morally wrong.[17] Thus, procedural moral norms that are outcome neutral appear to yield a picture of moral discourse that flirts with contradiction. In other words, it could seem that morality must be substantive (i.e., not at all procedural) "all the way down."

[17] The same is true if ϕ-ing is neither morally required nor wrong but rather permissible: the moral status of the action arguably depends on the relevant properties of ϕ-ing itself.

For example, suppose that a certain procedural *moral* norm directs hospital managers to normally settle ethical dilemmas—such as whether or not to approve a dangerous clinical trial—by following the advice of specially appointed ethics committees. Now, it seems that if the clinical trial is morally warranted, it must be so in virtue of the morally relevant properties of the trial itself (such as the extent of the risk to the subjects of the trial). This is so regardless of whether or not the ethics committee's advice is that the trial is or is not morally warranted. In contrast, the procedural moral norm for deciding the case prescribes following the committee's advice *regardless* of such morally relevant properties.

Assume that although the trial is morally unwarranted, the committee advises in favor of it. The dilemma of the hospital administrator, therefore, is choosing between following two conflicting moral norms. On the one hand, there is a substantive first-order norm:

N_1: The trial ought not to go forward.

On the other hand, there is a procedural second-order norm:

N_2: Decide whether N_1 according to the committee's advice.

Thus, the administrator appears caught on the horns of a moral dilemma.

Addressing this apparent paradox raised by the idea of a procedural moral norm, notice first that, *formally*, there is no contradiction between procedural and substantive moral norms. For instance, in the example of the clinical trial, norms N_1 and N_2 do not formally contradict each other. While N_1 prescribes *prohibiting* the clinical trial, N_2 prescribes *how to deliberate* on the moral status of the clinical trial.

Yet this is not enough to ensure the conceptual compatibility of a procedural moral norm with its relevant substantive moral norm. As a general matter, the absence of formal contradiction between two propositions does not immunize against other forms of conceptual defectiveness. Specifically, two moral norms, each prescribing an all-things-considered duty, cannot—as a matter of the nature of morality—conflict in the normative practical outcomes (that is, what you ought to do all things considered) of their prescriptions. For instance, if one is obligated to prohibit a clinical trial under N_1, then presumably N_2 is conceptually impossible, given that the normative practical outcome of following the prescription of N_2—requiring the clinical trial—is incompatible with the normative practical outcome of following the prescription of N_1—prohibiting the trial.

Dissolving this conceptual challenge to the idea of procedural moral norms requires rejecting the implicit premise that N_1 and N_2 prescribe *all-things-considered* duties as opposed to *pro tanto* duties (or even just *pro tanto* reasons). After all, there is nothing mysterious or problematic about conflicting *pro tanto* duties, and therefore, the occasional friction between substantive and procedural moral norms does not raise any special puzzle. Indeed, we believe that there can be distinct *pro tanto* first-order and second-order reasons, which can conflict with each other—and that the resolution of such conflicts is a moral matter, not a conceptual one.

Yet the interlocutor could object that this response to the conceptual challenge to procedural moral norms is too easy, as at least on the face of things, the paradox of moral procedural norms seems more intractable than a simple case of conflicting moral norms. This is because, seemingly, procedural and substantive norms can both provide *more* than merely *pro tanto* reasons. At least on the face of things, both substantive norms and procedural norms can "claim" categorical priority over one another, apparently leaving no space for balancing between them.

Nevertheless, we can explain away this sense of intractability by looking more closely at these two types of norms and their apparent incompatibility. On the one hand, as demonstrated above, substantive norms of the form "ϕ-ing is wrong" (e.g., "conducting the experiment is wrong") are naturally understood as supervening on all the morally relevant properties of the action ϕ itself, as opposed, for instance, to supervening on facts about how one ought to engage with the norm "ϕ-ing is wrong" (e.g., how one ought to deliberate on whether conducting the clinical trial is wrong). Therefore, substantive norms appear to provide all-things-considered reasons for or against ϕ-ing and are habitually assumed, in that sense, to be independent and categorically prior to procedural norms. That is, procedural norms arguably only bind if the action that they prescribe is itself morally permissible, regardless of the procedure.

On the other hand, given that procedural norms are second-order norms, namely, norms about (how to engage with) other (substantive) norms, balancing them against (let alone subordinating them to) the very same norms that they are about seems strange. This is perhaps most salient in the case of procedural norms of *forming and shaping* norms: it is strange to balance a procedural norm that governs the process of forming (substantive) norms against those very same (substantive) norms that are the outcome of that process. The same is true regarding procedural norms of *deliberation* bearing on the process of reasoning about other norms: seemingly, the norms governing the process of moral deliberation are not balanceable against the substantive norms that are the outcome of that same deliberative process.

Thus, these two types of norms appear incompatible, as they both appear to yield all-things-considered norms that are potentially conflicting. And discounting occasional moral tragedies or genuine moral dilemmas, moral discourse arguably cannot—as a conceptual matter—include conflicting all-things-considered norms.

In the example above, the hospital administrator is morally bound to adopt the advice of the expert ethics committee, following the procedural norm that "hospital administrators ought to decide ethical questions according to the advice of ethics committees." This procedural norm seems to stand even when following the committee's advice would yield an immoral outcome (in the eyes of the administrator). Indeed, administrators typically seem committed to the view that such a procedural norm is an all-things-considered moral edict. After all, were administrators to adopt committees' advice only when it coincided with their own views, administrators might as well just decide ethical matters on their own accord.

That said, in those cases wherein ethics committees' advice yields immoral outcomes, administrators, at least in their more reflective moments, do struggle with the unique moral dilemma between either: complying with the procedural moral principle that administrators ought to act in ethical matters according to expert advice, or adhering to the substantive moral norms governing the concrete case and their moral duty to administer their hospital ethically.

Still, allowing for a conceptual space for the coexistence of substantive norms and corresponding procedural norms requires discarding the intuitive assumption that both are all-things-considered norms. How then can we explain away the intuitive pull of this powerful (yet, we think, erroneous) assumption?

The answer, we hold, is found in the fact that substantive and procedural norms inhabit different orders. And, given this difference, it might seem natural to take both procedural and substantive moral norms as supreme over the other—even when that results in incompatible prescriptions. Indeed, the notion of balancing norms of different orders can even seem a misnomer, not unlike comparing apples and oranges. Thus, procedural and substantive norms that are supreme within their respective orders can understandably (yet still incorrectly) appear as all-things-considered norms that apply across all orders.[18]

[18] A norm is "supreme within an order" if it provides an overriding reason in respect of the specific action that that order is about. For example, in the case of the hospital administrator, there might be various considerations regarding what is the right procedure for deciding the substantive ethical question (whether or not the clinical trial is ethical). And while there are likely different *pro tanto* reasons in favor of certain procedures, there could also be a certain procedure that is the right procedure all things considered; namely, this would be the right procedure given all the (second-order) considerations in favor or against the available procedures. We stipulate that the ethics committee manifests such a procedure. In contrast, a norm is an all-things-considered norm "across all orders" if

This distinction between two types of supremacy helps to flesh out why procedural and substantive norms may mistakenly seem all-things-considered norms and why conflicts between them give rise to a novel type of moral dilemma (unlike standard conflicts between *pro tanto* normative considerations, such as utility and equality). Prescriptions of a norm that is supreme within one order are, in a sense, myopic as to the normative pull of reasons from other orders. After all, if you ask yourself the (second-order) question "What is the right procedure for deciding whether *ϕ-ing* is wrong?" in isolation (i.e., from first-order considerations), you are prone to conclude that you must (all things considered) follow the prescription of that appropriate procedure.[19] Returning to the hospital administrator's dilemma, it is the second-orderness of the procedural principle that administrators ought to manage their hospital according to expert advice that makes it appear as an all-things-considered norm functioning as a type of blinder, entirely filtering out from the administrator's deliberation those relevant moral norms not recognized by the ethics committee. Yet there is no *conceptual necessity* that the norm obligating administrators to follow the advice of ethics committees on certain matters of medical ethics is an all-things-considered norm across all orders. More generally, there is no conceptual constraint ruling out competition among norms across different orders. Accordingly, neither procedural moral norms (such as those discussed in section 4) nor the substantive moral norms that they are about are necessarily all-things-considered norms.[20] Therefore, appearances notwithstanding, it follows that there is no conceptual defect in the idea of a procedural moral norm.

3.2.2. *The Moral Objection*

Even accepting that procedural *moral* norms are conceptually sound, the interlocutor might still doubt such norms on moral grounds. Presumably, morally speaking, one ought not to follow a procedure that yields a prescription to do something that is in itself morally wrong. Thus, following our terminology, if it turns out that what we call "procedural moral norms" can contradict what

the action it prescribes is morally justified given all relevant reasons *regardless* of order. For example, if given all the relevant reasons from all orders, the morally right practical outcome is for the hospital administrator to decide as the committee advises, then N_2 is an "all-things-considered norm across all orders."

[19] If the interlocutor finds this analysis unnecessarily complicated, one can replace "all things considered across orders" with "all things considered" (*simpliciter*) and "supremacy [only] within a normative order" with "*pro tanto*," without altering our conclusion—which is, that procedural norms and substantive norms provide *pro tanto* reasons.

[20] Putting aside the possibility of unresolvable moral dilemmas or paradoxes.

we would label "substantive moral norms," then there appears reason to doubt that the former type of norms are indeed valid moral norms.

The remainder of the paper tackles this objection. Our argument in favor of the existence of procedural moral norms is twofold. First, we offer instances of what intuitively seem like genuine moral norms exhibiting the structure of what we characterize as a procedural norm. Second, space permitting, we shore up these intuitions by sketching possible lines of thought regarding the moral grounding of those procedural norms. In any case, our general point here is existential, demonstrating, in principle at least, that procedural moral norms of the kind described above exist. Hence, our argument that morality has a procedure does not depend so much on whether this or that specific procedural norm is justifiable.

4. RESPONDING TO THE MORAL OBJECTION: EXAMPLES OF PROCEDURAL MORAL NORMS

Above, deploying examples from law, we identified three broad types of procedural norms: norms of the application of norms; norms of deliberation on norms; and norms of forming, shaping, and validating norms. What follows are concrete examples of such types of norms, this time from within morality.

4.1. Procedural Moral Norms of Deliberation

Of the different types of procedural moral norms of deliberation, below we discuss two: norms of exclusion and epistemic norms.[21]

4.1.1. Exclusionary (or Discounting) Procedural Moral Norms of Deliberation

Let us begin with an example. Academic committees tend to believe that they ought to consider *only* certain reasons in favor of granting tenure. Such reasons include excellence in scholarship and teaching, administrative contribution, and collegiality. Let us call these reasons "academic." In fact, tenure committees typically tend to treat nonacademic reasons as falling outside the purview of their deliberation and official mandate, even if on balance those reasons morally outweigh the academic reasons. At the very least, academic committees tend to discount nonacademic reasons.

For instance, a candidate's emotional fragility is considered a peripheral or even illegitimate reason for a committee to grant him tenure. This is so even if assuming that on the balance of first-order moral reasons, the candidate's

[21] Other types possibly include, for example, procedural moral norms of deference, consultation, and exclusionary permissions. See Herstein, "Understanding Standing" and "Justifying Standing."

foreseeable suffering and overall well-being morally outweighs the moral value of the relevant academic reasons. Indeed, even if a candidate's well-being is a weighty moral reason, most academics are of the opinion that tenure committees ought to bracket such nonacademic reasons.

One explanation of this common practice of tenure committees is that the candidate's well-being is considered simply irrelevant to the committee's moral deliberations. Nevertheless, this point about "irrelevance" is patently erroneous. Clearly, as a moral matter, a person's well-being *is* morally relevant to decisions that may impact his well-being. In fact, on the balance of the relevant first-order moral reasons, avoiding derailing a person's life may often outweigh any negative or suboptimal academic impact that granting him tenure may bring.

A better account of this practice of tenure committees ignoring nonacademic reasons views it as following an exclusionary norm. An exclusionary norm is a norm directing agents to ignore or not act on certain other norms or reasons.[22] Exclusionary norms are second-order norms because they are about other norms—namely, they direct one to exclude those other norms. In our example, such an exclusionary norm directs tenure committees to exclude nonacademic norms and reasons from deliberation.[23]

But is this exclusionary norm *procedural*? As argued at the outset, second-order norms are not necessarily procedural. Arguably, the same is true of exclusionary norms, which are a type of second-order norm. A *substantive* exclusionary norm may, for instance, direct that when confronted with two competing first-order norms, one ought to prioritize one norm in favor of the other. For example, assume that parents have both a reason to provide their child with food that the child finds tasty—such as a hamburger—as well as a reason not to partake in the exploitation of animals. Some believe that in this type of clash, considerations of one's child's culinary delights ought to be excluded entirely rather than weighted against considerations of animal rights.[24] Such a second-order norm is substantive rather than procedural, given that it bears on the normative relationship between the two first-order norms—for example, assigning lexical priority to one norm over the other—and not on the *process of how* to engage with those norms.

We argue, however, that some instances of exclusionary norms are only defendable if taken as procedural rather than substantive. Such is the case in our

[22] Raz, *Practical Reasons and Norms*, 35–48.

[23] Enoch explores such exclusionary norms under the heading of "quasi-protected reasons" ("Authority and Reason-Giving," 321).

[24] Adams, "In Defense of Exclusionary Reasons."

example of the tenure committee, which is crucially different from the norm we discussed in the aforementioned example of the hamburger.

Notice first that the exclusionary norm guiding the tenure committee bears the hallmarks of a procedural norm. Recall that a second-order norm is procedural if it bears on *how to engage* with other norms. That is, as explained, procedural norms are in a sense outcome neutral—they are about the process of such engagement as opposed to the outcome of the engagement itself. The case of the tenure committee exhibits these procedural hallmarks: when excluding nonacademic reasons from deliberation, tenure committees are acting on a norm that does *not* bear on the matter of whether or not to grant tenure, unlike the example of the hamburger, where the exclusionary norm assigned (lexical) normative priority to one norm over the other, in the case of the tenure committee, the norm is to simply ignore one norm *despite* its relevance and apparent greater moral weight. In that sense, this exclusionary norm is outcome neutral and, therefore, procedural.

But what can justify excluding relevant and even weighty reasons from deliberation? Presumably, all reasons bearing on a practical matter ought to be part of the practical deliberation pertaining to that matter. What, in other words, makes such a procedural norm moral? In the context of our example, one path toward understanding the justification for such exclusionary norms begins with observing certain similarities they share with what is known as "role morality." As T. M. Scanlon puts it, being a good teacher, for instance, involves bracketing and reordering the reason-giving force of some norms that otherwise might be quite relevant.[25] An exclusionary procedural norm of deliberation and the idea of role morality thus share a key feature—namely, bracketing relevant reasons. This suggests that perhaps certain procedural norms and certain instances of role morality stem from the same moral grounds. Moreover, viewing role morality through a procedural prism suggests that it has an overlooked procedural dimension.

A plausible account of role morality is that its justification is tied up with the value of the relevant social institution the role is couched in. For instance, returning to our example, perhaps the role morality of academics derives from the value of the institution of academia. But why does realizing the value of academia require the exclusion of relevant nonacademic moral reasons in the workings of tenure committees? After all, presumably, were a tenure committee to promote a mediocre candidate on the grounds of nonacademic reasons, the institution of academia would not as a whole suffer any significant setback.

[25] Scanlon, *What We Owe to Each Other*, 52.

Notice first that if taken as a *substantive* exclusionary norm, such a norm seems morally invalid. For instance, suppose that the tenure committee's members took themselves to be subject to a substantive norm of exclusion, assigning lexical priority to academic reasons over nonacademic reasons. It would then follow that such committees were following a morally invalid norm, because as we already argued above, in each particular case, derailing a person's life is *not* lexically inferior to the utterly negligible disvalue that adding yet another mediocre academic would have for the institution of academia. Accordingly, understanding such exclusionary norms as substantive implies that they are invalid. Thus, if such exclusionary norms are possible, they are only so if they are procedural.

Beyond this transcendental argument, notice that the view that the moral norms at hand are procedural is plausibly supported across different approaches to normative ethics. One justification for such exclusionary norms has a rule-consequentialist flavor. Although in every particular case a tenure committee morally ought to consider nonacademic reasons, over time, in aggregate, the overall academic quality would greatly erode. Such a rule-consequentialist justification of exclusionary norms only justifies them as *procedural* norms of exclusion, because were tenure committees allowed to deliberate substantively in the case of every subpar candidate, then in each particular case the appropriate moral decision would be to grant tenure. This is because every *particular* case of granting tenure to an undeserving candidate is negligible in terms of eroding the overall quality of academia, not justifying the harm to the individual candidate, and the committee's granting of tenure to an undeserving candidate does not make it more likely that other tenure committees would do the same. Thus, morally correct *substantive* reasoning leads to morally suboptimal results—calling for a procedural norm blocking this type of reasoning.

On this account, then, such exclusionary norms are not justified by the properties of the action itself. Indeed, were the agent to apply good moral reasoning in deciding whether or not to perform an action, her inevitable conclusion must be that she ought to violate the exclusionary norm and act according to all the moral reasons relevant to the action—which is not what morality "wants." It seems, therefore, that in cases such as the tenure committee example, the justification for the exclusionary rule is, in a sense, normatively inaccessible to the agent who is occupying the first-person perspective. Our point is not that the agent cannot reflect on the dilemma from the third-person perspective or that she is unable to comprehend the reasons in favor of the general exclusionary rule. Rather, even in the face of such awareness, when acting from the first-person point of view in a particular case, it seems that one cannot morally dislodge oneself from that point of view such that one ought not to obey such procedural norms. More generally, if we are correct about the existence of such procedural moral norms, it follows that the moral status of actions turns on more than just the properties of the action itself, such as the action's consequences (both direct and indirect), the agent's intentions and other mental states, and so on. Accordingly, procedural exclusionary norms impact the moral status of actions in ways that are, in that sense, inaccessible from the first-person point of view of the moral agent.

Such procedural exclusory norms are also plausibly justifiable within other paradigms of moral theory, such as approaches with a more Kantian flavor. What we have in mind here is the conviction that agents ought to rely on an impartial and general point of view to morally guide their actions. Returning to the example of the tenure committee, including morally relevant nonacademic reasons in their deliberation is only morally allowed if the members of such committees can at the same time will that it become a "universal law" governing all tenure committees. That is, tenure committees ought to act impartially, treating the specific tenure candidate before them as they believe any and all such candidates ought to be treated under similar conditions. And presumably, such an impersonal and general view could mandate excluding nonacademic yet morally relevant reasons from the committee's deliberations on whether to grant tenure.

Procedural exclusionary moral norms may also find grounding in approaches to normative ethics focused on values that seem neither rule-consequentialist nor Kantian. Consider Bernard Williams's position that deep personal attachments can permit and even mandate certain actions. In Williams's well-known example, a man is faced with the dilemma of choosing to save his wife or a stranger. Most people intuitively believe that the man is permitted or even obliged to act partially toward his wife. Williams points out the inadequacy of explaining this intuition from an impartial paradigm in morality, a criticism applicable both to a Kantian and a rule utilitarian. For Williams, the man is permitted (or even obliged) to save his wife over the other person for the straightforward reason that *she is his wife*. Were we to ground this permission in further moral reasons—such as that it would bolster the institution of marriage (a rule-utilitarian reason) or some universal permission or duty to prioritize one's spouse (a quasi-Kantian reason)—it would fail to express and even offend against the man's deep personal attachments. For Williams, this would count as "one thought too many."[26]

Under an approach such as Williams's, what grounds the exclusion of morally relevant reasons is that their *mere consideration* is objectionable. This

[26] Williams, *Moral Luck*, 18.

makes Williams's exclusionary norm procedural, as it is directed at the process of engaging with reasons and norms as opposed to their substantive weighing. Such procedural norms embody a principle of what may be coined "moral inadmissibility"—the very admittance of the reason into one's moral reasoning is what is morally deficient—reminiscent of the legal inadmissibility of probative yet illegally obtained evidence, the very admittance of which mars the integrity of the judicial process.

One may object to our notion of procedural norms by claiming that it collapses into norms of role morality. Relatedly, norms of role morality appear analogous to norms of legal procedure, as the latter are directed toward legal officials fulfilling various roles, such as judges and legislators. If true, this would suggest that procedural norms are always role-oriented norms, making our project of unearthing such norms in morality somewhat trivial or even merely terminological.[27]

In response, while maintaining a measure of overlap, it is the case neither that all norms of role morality are procedural nor that all procedural norms are part of role morality. For example, the norm that "members of tenure committees *qua* members ought to grant tenure based on the candidates' academic record" is arguably one of role morality, yet it is substantive rather than procedural. Conversely, the categories of procedural moral norms detailed below contain many instances of norms that are not part of role morality.

4.1.2. Procedural Epistemic Moral Norms

Presumably, in moral matters we have a moral reason to act on the best available evidence.[28] And such moral reasons may ground certain *epistemic* procedural moral norms of deliberation.[29] In fact, we have already encountered such a norm in the example of the medical ethics committee. One plausible justification for the second-order norm to follow the advice of the ethics committee is the committee's superior medical and putative moral expertise,[30] providing

[27] For objections along these lines, we are grateful to the anonymous reviewer for this journal.

[28] Incidentally, this reason is procedural. It is about how one ought to decide how one ought to act—namely, on the best available evidence. And it is outcome neutral, as it bears not on how one ought to act but on how one ought to determine how one ought to act.

[29] "Epistemic moral norms" are distinct from "epistemic norms." While the latter are about what one ought to believe, the former are about how one ought to act given certain epistemic conditions.

[30] The justification of a norm prescribing the hospital manager to follow the advice of the ethics committee might rely on the epistemic advantages of the particular committee or on rule-consequentialist grounds: that following the advice of ethics committees brings about morally better consequences in general, even if in some particular cases hospital managers relying on their own moral judgment yields morally better outcomes.

the best available *evidence* on the matter.[31] Another example of a procedural epistemic moral norm is drawn from the debate on procedural rights. Enoch has recently defended procedural moral rights, offering the following example in response to Wellman's view that pre-institutional procedural rights do not exist: a colleague spreads malevolent rumors about you at work.[32] While you do not know which of your colleagues it was, you randomly pick one of them, blame him, resent him, and stop inviting him to lunch. Providentially, it turns out that it was indeed he who had spread the rumors. Yet notwithstanding the colleague's blameworthiness, Enoch argues that in blaming him without sufficient evidence you still wrong him (though, perhaps, he is not in a position to complain about your treatment of him).

Indeed, it makes perfect sense that morality would contain norms about how agents ought to act when faced with epistemic uncertainty. Presumably, it is agents' epistemic shortcomings that invite procedure into the moral discourse, as we are regularly (perhaps always) called upon to decide how we ought to act under conditions of evidentiary or other epistemic imperfection. In fact, in an epistemically ideal world, where all moral truths and all morally relevant factual truths are readily known to the agent, there appears to be no room for epistemic moral norms. More generally, arguably, one explanation of the existence of procedural norms in morality is that morality is for agents who are by their very nature imperfect. Thus, in epistemic procedural moral norms (such as in the examples above) the moral status of an agent's action may depend on her mental state and not only on the objective state of the external world.

This realization invites the objection that the distinction between procedural and substantive norms collapses into the familiar distinction between objective and subjective oughts. An objectivist view about norms (e.g., oughts, duties, and rights) is that what determines what one ought to do is the objective state of affairs of the world, whereas under a subjectivist view what matters are one's *beliefs* about that state of affairs. For instance, in the case of the slandering colleague, subjectivism implies that whether you violated your colleague's rights turns not on whether he is in fact guilty of spreading the rumors but rather on whether or not you believe he is guilty. Of course, assuming subjectivism about ought, randomly picking the colleague to be blamed violates his rights even if he is in fact guilty. Thus, the objection is that procedural moral norms

[31] While the category of epistemic procedural moral norms (on which we focus here) is underexplored, there are discussions of moral norms prescribing epistemic procedures. See, e.g., Rosen, "Skepticism about Moral Responsibility," 301.

[32] Enoch, "In Defense of Procedural Rights (or Anyway, Procedural Duties)"; and Wellman, "Procedural Rights."

exist only if we assume subjectivism about ought, but then, the objection continues, procedural moral norms are simply substantive norms with subjective oughts.[33] For example, if the "ought" in the procedural norm "you ought not to blame your colleague without sufficient evidence" must be a subjective ought, then this norm seems to have the same meaning as the norm "you ought not to blame your colleague" (with the "ought" understood as subjective), which is arguably a *substantive* norm only with a subjective ought. Indeed, it might ostensibly appear that the nature of the "ought" in procedural epistemic moral norms (such as "one ought to make moral decisions based on good evidence") must be subjective. This is wrong, however, as becomes apparent once we formulate the relevant procedural norm as a fully fleshed out procedural norm, that is, a second-order norm bearing on the engagement with another norm.[34]

For instance, in the case of the slandering colleague, the relevant procedural norm is "you *ought not* to decide whether you are morally *permitted* to blame or punish your colleague without sufficient evidence." In this example, each of the two normative terms found in the procedural norm may take either an objective or a subjective form. Accordingly, there are four possible combinations of such normative terms in procedural norms: the "procedural ought"—namely, the external ought referring to the decision-making—may be objective or subjective, and the same is true of the internal normative term "permitted" embedded in the scope of the procedural (external) ought. Of the four combinations, what we wish to stress is that the pairing of two *objective* normative terms is a possible and even plausible account of many procedural moral norms. For example, the normative terms in the norm "you *ought not* to decide whether you are morally *permitted* to blame or punish your colleague without sufficient evidence" are plausibly both objective: the internal normative predicate ("morally permitted") can be objective because the aim of your moral deliberation (determining whether you are permitted to blame your colleague) is finding the objectively correct moral answer (is he blameworthy for spreading the rumors?); the procedural (external) ought is *also* plausibly objective, for it plausibly prescribes deciding based on the best available evidence, not what you *believe* is your best available evidence.

To conclude, looking at these intuitive procedural epistemic moral norms, we learned that procedural moral norms are compatible with objectivism about

[33] For a similar objection—raised against the idea of pre-institutional procedural rights—see Wellman, "Procedural Rights."

[34] For different responses to Wellman, see Enoch, "In Defense of Procedural Rights (or Anyway, Procedural Duties)"; Adams, "Grounding Procedural Rights"; and Stewart, "Procedural Rights and Factual Accuracy."

"ought" and, therefore, the objection that procedural moral norms are in fact substantive norms with subjective oughts fails.

4.2. Procedural Moral Norms of the Application of Norms

Moving on from norms of deliberation, another type of procedural moral norm bears on the manner and means of *applying* another norm, thereby governing the practical aspects of engaging with such other norms. As we saw, in the law, the proliferation of such procedural norms seems almost trivial. Our view is that morality too contains such norms.

Consider, for example, an editor of an academic journal charged with deciding whether or not to accept a submission written by her PhD student. Presumably, the editor is subject to a moral norm according to which editors ought not to decide whether a paper authored by someone close to them ought to be accepted. This norm is a procedural moral norm of application: it is second-order, as it is about applying the norms determinative of academic quality; it is about the "how" of engaging with those norms, namely, determining when one ought to disqualify oneself from deciding how to apply those norms; and it is outcome neutral, given that it is agnostic as to whether or not the submission is worthy of acceptance. Such procedural moral norms for avoiding conflicts of interest can be justified on epistemic grounds. Yet even if making such normative judgments under a conflict of interest does not pose any epistemic deficiency, doing so still seems morally problematic.

Another example is moral norms of hearings. In law, norms mandating hearings are plentiful and uncontroversial, appearing at least partially grounded in counterpart moral norms of hearings. Indeed, like legal principles against conflicts of interest, legal norms of hearings are labeled principles of "natural justice," suggesting that the law in a sense transplanted them directly from pre-institutional morality.[35]

To demonstrate that procedural norms are not exclusively institutional, here is an example of a norm of hearing from a purely interpersonal context. Consider the following two norms. "One is at liberty to sever a romantic relationship" seems like a morally sound, first-order norm. And the following norm also seems sound: "at least sometimes, one ought not to exercise one's moral permission to sever a romantic relationship without first hearing one's lover's response prior to the separation."[36]

[35] Shauer, "English Natural Justice and American Due Process."

[36] To be clear, we do not claim that such norms obtain when exiting all romantic relationships, such as in the case of abusive relationships.

The latter norm is procedural. First, it is second order—the thing the normative status of which is at stake (exercising "one's moral permission to sever a romantic relationship without first hearing one's lover's response prior to the separation") is defined in moral terms ("exercising one's moral permission"). Here the second-order norm is about the manner and means of applying the former (first-order) norm, namely, prior to acting on the norm permitting breaking up with one's lover, one must grant her a hearing. Second, this norm is about the "how" of engaging with another norm. Namely, it sets conditions for how to apply the norm permitting one to sever a relationship, mandating granting a hearing prior to exercising one's permission to separate. Finally, it is outcome neutral—the norm does not directly bear on whether one ought to exercise one's permission to separate, only on how one ought to do so.

Now, hearings mainly have two normative roles. First, hearings typically fulfill an epistemic function. In our example, a hearing potentially provides relevant evidence of one's reasons for the breakup. In that sense, norms mandating hearings are procedural (epistemic) norms of deliberation similar to those discussed above. Second, seemingly, hearings also carry normative significance not reducible to their epistemic virtues. In our example, we can stipulate that one unequivocally knows all the relevant facts for the decision to separate, and *still* she ought to hear out her lover prior to breaking up with her. Indeed, nonepistemic norms demanding hearings are ancient, presumably present even at the genesis of humanity, as God himself—who is presumably omniscient—provided Adam and Eve with a hearing, allowing them to confess and explain their sins before he banished them from the Garden of Eden.[37]

Although intuitive, explaining the nonepistemic moral significance of hearings is not trivial, be it in law or in our more interpersonal context.[38] Here we can only gesture toward a justification, such that when severing a meaningful relationship, delaying one's final decision until after a hearing expresses respect toward one's lover and what they shared. Such a norm prescribing a delay to the decision to break up would be outcome neutral and relevantly similar to *legal* norms of hearing. A different norm may prescribe no delay to one's final decision to break up but a delay to the act of breaking up itself (or of announcing the decision) until after a "hearing." Possible justifications include softening the blow of the breakup, providing the opportunity for "closure," and acting in a manner that is less cold and humiliating. While such delays are sometimes required, one can object that a norm prescribing giving one's spouse her say prior to effecting the breakup but after having made the final decision to do so

[37] Gen. 3:8–13.

[38] In favor of hearings in public law, see, e.g., Harel, *Why Law Matters*.

is a *substantive* (second-order) norm. This is because it does not bear on the process of one's engagement with another norm and is not outcome neutral but rather provides a direct prescription regarding the final action (the breaking up). Yet while we acknowledge that there might be norms that are not easily classifiable as procedural or substantive, this norm *is* plausibly conceived as procedural, for it does not bear directly on the content or scope of the right to break up but only on the manner and means of how this right ought to be exercised.[39]

4.3. Procedural Moral Norms of Forming, Shaping, and Validating Norms

While it is perhaps less surprising to encounter procedural *moral* norms of application and deliberation, such as natural justice norms or the epistemic norms explored above, the notion that morality involves norms bearing on the procedure of forming, shaping, and validating other moral norms might seem especially puzzling. After all, there are no "parliaments of morality" engaged in the formation of moral norms. Likewise, unlike legal norms, the validity of moral norms seems independent of meeting procedural conditions, certainly when it comes to mundane procedures such as convening the requisite quorum in the House of Representatives.

Nevertheless, our view is that morality does involve such procedural norms. Notice that we do not argue that *all* substantive moral norms are products of procedures for the forming and shaping of moral norms. In fact, it seems to us plausible that there are basic moral norms the content and validity of which are independent of any procedure whatsoever. That said, there are of course those who do believe that the ultimate criteria for moral validity or for political morality are at their core, procedural and outcome neutral.[40] As already noted, we do not delve into this philosophical divide, as our aim is to offer examples of procedural moral norms of forming, shaping, and validating norms agreeable to both metaethical camps.

Suppose that the founders of a philanthropic foundation are considering investing the foundation's resources in one of two worthy causes: reducing

[39] In that sense, this norm is different from second-order substantive norms such as the norm "punishing a person for an action that is morally permissible is morally wrong," which bears directly on the moral permissibility of the punishing itself rather than merely on the *process* of one engagement with any other norm (see section 1.3).

[40] E.g., Rawls derives his two principles of justice by employing a hypothetical choice procedure (the "original position") designed to "incorporate pure procedural justice at the highest level" (*A Theory of Justice*); see also Rawls, "Justice as Fairness." Some even hold that moral validity in general is dependent on certain procedures (see, e.g., Korsgaard, "The Normative Question").

extreme poverty or curing infectious diseases. Let us assume for the sake of argument that reducing extreme poverty is the morally right choice under these circumstances. However, the founders cannot reach a consensus on which goal the foundation ought to adopt. To break the deadlock, the founders opt to resolve their disagreement by voting, and the majority votes that the foundation ought to adopt curing infectious diseases as its morally superior goal. Suppose that you are the foundation's CEO. You now face a dilemma between following one of two norms: the first-order, substantive norm, "the foundation ought to support reducing extreme poverty"; and the second-order, procedural norm, "the foundation *ought* to follow the results of the vote on whether the foundation *ought* to support reducing extreme poverty or curing infectious diseases."[41]

This moral dilemma exposes the curious tension we already encountered between procedural and substantive norms. Prior to the vote, you ought to have pursued a policy of reducing extreme poverty. Yet after the vote, what you ought to do is less clear. At the very least, you have some moral reason to comply with the procedural norm: that is, the vote created a new moral reason for favoring curing infectious diseases over reducing extreme poverty. Accordingly, the procedural norm at play prescribes complying with the results of the vote regarding what the foundation ought to do, functioning as a procedural moral norm for forming new norms.

A possible objection to our formulation of the above procedural norm as second-order is that it is artificial, overly stretching natural language. After all, why not articulate this norm as "you ought to follow the results of the vote on what *to do*" rather than as "you ought to follow the results of the vote on what you *ought to do*." Under this objection, the founders' vote is not normative; that is, it is a vote about what the foundation *will* do, not about what it *ought* to do. And if so, the procedural norm prescribing following the results of the vote involves only one "ought," and therefore there is no second-order norm at play here, thus dissolving the tension we alluded to above.

However, although our articulation of the norm as second-order may appear artificial, we stand by it. First, it is true that the founders' vote yields a practical decision and is not a vote on a theoretical question (such as the questions discussed in an academic course on normative ethics).[42] Yet the founders

[41] Further examples of voting procedural norms are plentiful in morality. Such norms can be noninstitutional and are not necessarily instances of role morality. For example, children regularly decide what game they ought to play by majority vote.

[42] A procedural norm prescribing voting for resolving a purely *theoretical* moral disagreement (as opposed to a practical one) is an epistemic norm about what one is justified in believing rather than a moral norm about what one ought to do.

are engaged in a moral deliberation exactly about what the foundation *ought* to do, and the vote is the procedure for settling *that* question. Moreover, the vote is a process for settling a moral disagreement on a practical question *normatively*. As such, voting is morally laden. It is not like settling a moral conflict by rumble. As discussed above (section 1.4.), embedded in a vote on "what to do" is a determination of what "one ought to do" by the very fact that the matter is settled by vote. Accordingly, there is a procedural norm at play here.

Moral justifications in favor of voting as a process for collective moral decision-making are many. Briefly zeroing in just on our example, voting seems a morally sound process for deciding what the foundation ought to do. Obviously, there are instrumental justifications for voting, such as securing coordination among the founders and assuring that the philanthropic venture at least gets off the ground. Another possible justification is epistemic, at least if assuming that the founders are epistemic peers, and in the absence of other superior experts on the matter, voting indeed seems epistemically rational and for that reason morally justified.[43] Whether noninstrumental justifications also obtain in our case is less obvious, because unlike the context of democracy, where typically the moral patients of the vote largely overlap with the electorate, in our case there is a complete separation between the voters and those whom the vote impacts. Thus, justifications from consent, liberty, fairness, equality, and membership appear less fitting.

Other examples of forming procedural moral norms are norms prescribing conducting lotteries as a way of settling questions of allocation.[44] For example, suppose *A* and *B* are equally deserving claimants to a certain indivisible good. Many believe that in such a case one ought to allocate the good by lottery. The moral norm prescribing this process of allocation cannot be first-order—for as stipulated, there is no moral reason to prefer the claim of one claimant over the other's. Rather, the moral norm prescribing the lottery is necessarily a second-order norm along the lines of something like "one *ought* to conduct a lottery to determine how one ought to allocate an indivisible good between equally entitled claimants *A* and *B*."

Now, suppose that *A* wins the lottery. While prior to the lottery there was no reason to prefer *A* over *B*, after the lottery there is such a reason. The lottery therefore forms a new moral reason. And, accordingly, the second-order

[43] For epistemic justifications of voting and democratic procedures, see, e.g., List and Goodin, "Epistemic Democracy"; and, Estlund, *Democratic Authority*.

[44] In fact, Rawls viewed lotteries as the paradigmatic case of what he calls "pure *procedural* justice"—namely, the case where "there is no independent criterion for the right result: instead there is a correct or fair procedure such that the outcome is likewise correct or fair" (*A Theory of Justice*, 75).

norm prescribing *how* to create such a reason—namely, by the procedure of lottery—is a norm-forming procedural norm. Relatedly, further reflecting its procedural nature, this norm is outcome neutral in the sense that it bears on the *process* of forming the first-order norm ("you ought to allocate the good to *A*"), *not the content* of that norm. The second-order norm is indeed agnostic as to whether it is *A* or *B* who should receive the good. It determines this normative outcome only indirectly.[45]

5. CONCLUSION

Unlike law, morality is arguably neither posited nor institutional. And still, much like law, morality not only prescribes various procedures (which seems uncontroversial) but also contains norms that are themselves procedural. Although the coexistence of procedural norms alongside substantive norms might at first blush seem paradoxical, we argued both that the idea of procedural moral norms is conceptually sound and that some such norms are morally grounded. In this respect, morality is not substantive through and through and, therefore, is more like law than what one might have expected.[46]

Hebrew University
King's College London
ori.herstein@mail.huji.ac.il

Hebrew University
malcai@mail.huji.ac.il

REFERENCES

Adams, N. P. "Grounding Procedural Rights." *Legal Theory* 25, no. 1 (January

[45] Justifications for lotteries vary. One common justification of allocation by lottery is grounded in assuring impartiality. For instance, Peter Stone (in *The Luck of the Draw*) argues that impartiality animates a second-order norm (which he calls a "meta-principle") justifying those lotteries that sanitize distributions from morally irrelevant or corrupting considerations.

[46] For helpful written comments and discussions, we are grateful to Ilan Benshalom, David Enoch, Miriam Gur-Arye, Alon Harel, Andrei Marmor, Yifat Naftali, Joseph Raz, Massimo Renzo, Eyal Zamir, Lorenzo Zucca, the participants of the King's College London's Legal Philosophy Workshop (London, 2020), the Varieties of Moral Address Workshop (Jerusalem, 2021), the Hebrew University Law Faculty Seminar (Jerusalem, 2021), and the referees for this journal.

2019): 3–25.

———. "In Defense of Exclusionary Reasons." *Philosophical Studies* 178, no. 1 (January 2021): 235–53.

Alexander, Larry. "Are Procedural Rights Derivative Substantive Rights?" *Law and Philosophy* 17, no. 1 (January 1988): 19–42.

Bagnoli, Carla. "Constructivism in Metaethics." In *Stanford Encyclopedia of Philosophy* (Spring 2021). https://plato.stanford.edu/entries/constructivism-metaethics.

Enoch, David. "Authority and Reason-Giving." *Philosophy and Phenomenological Research* 89, no. 2 (July 2014): 296–332.

———. "Can There Be a Global, Interesting, Coherent Constructivism about Practical Reason?" *Philosophical Explorations* 12, no. 3 (July 2009) 319–39.

———. "In Defense of Procedural Rights (or Anyway, Procedural Duties): A Response to Wellman." *Legal Theory* 24, no. 1 (April 2018): 40–49.

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

Harel, Alon. *Why Law Matters.* Oxford: Oxford University Press, 2014.

Hart, H. L. A. *The Concept of Law*. Oxford: Oxford University Press, 1961.

Herstein, Ori J. "Justifying Standing: Hypocrisy, Minding Your Own Business, and Knowing One's Place." *Philosophers' Imprint* 20, no. 7 (April 2020): 1–18.

———. "Understanding Standing: Permission to Deflect Reasons" *Philosophical Studies* 174, no. 12 (December 2017): 3109–32.

Korsgaard, Christine M. "The Normative Question." In *The Sources of Normativity*, 7–48. Cambridge: Cambridge University Press, 1996.

List, Christian, and Robert E. Goodin. "Epistemic Democracy: Generalizing the Condorcet Jury Theorem." *Journal of Political Philosophy* 9, no. 3 (December 2001) 277–306.

Malcai, Ofer. "Second-Order Propositions and Metaethical Neutrality." Unpublished manuscript.

Malcai, Ofer, and Ronit Levine-Schnur. "Which Came First, the Procedure or the Substance? A Few Notions of Priority." *Oxford Journal of Legal Studies* 34, no. 1 (Spring 2014): 1–14.

O'Neill, Onora. *Acting on Principle: An Essay on Kantian Ethics*. 2nd ed. Cambridge: Cambridge University Press, 2013.

Rawls, John. "Justice as Fairness: Political Not Metaphysical." *Philosophy and Public Affairs* 14, no. 3 (Summer 1985): 223–51.

———. *Lectures on the History of Moral Philosophy.* Edited by Barbara Herman. Cambridge, MA: Harvard University Press, 2000.

———. *A Theory of Justice.* Cambridge, MA: Harvard University Press, 1971.

Raz, Joseph. *Practical Reason and Norms*. 2nd ed. Oxford: Oxford University

Press, 1990.

Rosen, Gideon. "Skepticism about Moral Responsibility." *Philosophical Perspectives* 18 (2004): 291–304.

Rosenthal, Chelsea. "What Decision Theory Can't Tell Us about Moral Uncertainty." *Philosophical Studies* 178, no. 10 (October 2021): 3085–105.

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

Shauer, Frederick F. "English Natural Justice and American Due Process: An Analytical Comparison." *William and Mary Law Review* 18, no. 1 (1976): 47–72.

Solum, Lawrence B. "Procedural Justice." *Southern California Law Review* 78. no. 1 (November 2004): 181–322.

Stewart, Hamish. "Procedural Rights and Factual Accuracy." *Legal Theory* 26, no. 2 (October 2020): 156–79.

Stone, Peter. *The Luck of the Draw: The Role of Lotteries in Decision Making*. Oxford: Oxford University Press, 2011.

Wellman, Christopher. "Procedural Rights." *Legal Theory* 20, no. 4 (March 2015): 286–306.

Williams, Bernard. *Moral Luck.* Cambridge: Cambridge University Press, 1981.