JOURNAL *of* ETHICS
& SOCIAL PHILOSOPHY

VOLUME XXVI · NUMBER 1
November 2023

ARTICLES

DISCUSSIONS

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

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

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

Journal of Ethics and Social Philosophy
vol. 26, no. 1 · november 2023

https://doi.org/10.26556/jesp.v26i1.2462

BACKSLIDING AND BAD FAITH

ASPIRATION, DISAVOWAL, AND (RESIDUAL) PRACTICAL IDENTITIES

Justin F. White

> For too much of my Life,
> I've apologized when I wasn't wrong,
> all to make a situation better.
> I'm not going to be that person anymore.
> —Samantha King, *Born to Love, Cursed to Feel*

In one way or another, we have probably all been Samantha. We have seen something in ourselves that we dislike, and we have committed to change. The now undesired tendency could have been consciously cultivated—maybe Samantha decided to start apologizing as a way to keep peace. Or it could have been somewhat passively acquired, maybe as a coping mechanism to defuse tense situations. But whatever the tendency's origins, when Samantha says, "I'm not going to be that person anymore," she is committing to change—perhaps to hold her ground when her position is justified and not to apologize merely to avoid unpleasantness. If Samantha is like most of us, however, despite her sincere aspiration to change, the odds are that she will eventually (maybe repeatedly) find herself apologizing when tensions rise, even when she is not wrong. But when Samantha falls into that unwanted habit and apologizes, what sort of person is she? Is she still, or again, "that person," despite her commitment to change? Or does her commitment itself change the sort of person she is, even when she backslides? And if so, how?

On the one hand, one could argue that our actions are the best indicators of the sorts of people we are. As Inez says in Sartre's *No Exit*, "It's what one does, and nothing else, that shows the stuff one's made of."[1] According to this view, Samantha's conflict-avoidant apologies show who she is better than her stated desire or commitment to change: if she apologizes when she is not wrong, she is still (or is once again) "that person." On the other hand, not everything we

[1] Sartre, *No Exit*, 43.

do reflects who we are in the same way. Some apologies are unprompted and wholehearted. Others are begrudging, perhaps coming after significant prodding. That Samantha would be frustrated with and disappointed in herself for an unjustified apology, but happy if she were to stand her ground, suggests that the verbal apology ("what one does," in Inez's formulation) is only part of the story. And the details of the story matter.

Sometimes we explicitly say that an action reflects our core values or ideals. Perhaps more familiar is the way we sometimes seek leniency for poor behavior by claiming that the bad behavior is an aberration. We might say, "That's not who I really am," in order to distance ourselves from actions and perhaps to signal that we are trying to change, that the actions no longer represent our core values. But we can say the same words when we are simply in denial. And on the face of it, aspiring is different from being in denial. And different responses seem appropriate for backsliding aspirants like Samantha than for those who consistently seek what Harry Frankfurt calls "unmerited indulgence," who seek "forgiveness" but have no interest in changing bad behavior.[2] But how, exactly, are they different? And what do these differences suggest about the nature of the self and our relation to our actions?

To think through these questions, I use the notion of a *practical identity*, which Christine Korsgaard defines as "a description under which you value yourself." [3] The term has been widely adopted, even if the details vary and are sometimes unspecified.[4] Broadly speaking, practical identity refers to whatever forms one's practical outlook. Common examples are characteristics or roles, such as *parent*, *lover*, *teacher*, or, in Samantha's case, someone who does not apologize when she is not wrong. Admittedly, practical identity is a wriggly notion. It can be hard to pin down because we usually have various roles and characteristics. So, depending on how one parses it, one typically either has a complex practical identity or multiple identities. Korsgaard's account seems to be that we do have various particular practical identities—such as student, parent, or lover—but that part of the task of self-constitution is to integrate those roles into a single identity.[5] But practical identity is also wriggly because although

[2] Frankfurt, "Identification and Externality," 63.

[3] Korsgaard, *Sources of Normativity*, 101.

[4] For a sample of views tacitly assuming, explicitly referencing, or critically engaging with Korsgaard's notion of practical identity, see Velleman, "Willing the Law"; Crowell, "*Sorge* or *Selbstbewußtsein*?" and "The Existential Sources of Normativity"; Atkins and McKenzie, *Practical Identity and Narrative Agency*; Lear, *A Case for Irony*; Wallace, *The View from Here*; Shoemaker, *Responsibility from the Margins*; Westlund, "Who Do We Think We Are?"

[5] In *Sources of Normativity*, she writes, "Practical identity is a complex matter and for the average person there will be a jumble of such conceptions" (101). With her example of a

our self-conceptions matter, as Korsgaard's account emphasizes, our practical identities seem to outstrip our self-conceptions. We often see and engage with the world in ways that go beyond and can be at odds with our self-conceptions.

To explore these issues, we will discuss several fictional or fictionalized examples. In addition to Samantha, we will talk about a jealous individual and a parent who is trying to find better balance between work and family. To develop a multidimensional account of practical identity that captures the nuances of these cases, I use Korsgaard's account and Steven Crowell's Heideggerian alternative to distinguish between two dimensions of practical identity: *reflective practical identity* (roughly: how we see or think of ourselves) and *phenomenological practical identity* (roughly: the broader selves that shape how the world appears to us).[6] These dimensions typically involve feedback loops and so typically coordinate with each other in various ways. When I have the reflective practical identity of parent, for example—thinking of myself as a parent—there is coordination with my phenomenological practical identity when parental possibilities are salient. And in the other direction, when I inhabit the world of a parent—seeing the world in terms of parental possibilities—I am more likely to think of myself as a parent. Most of the time, then, because of this coordination, how we think of ourselves affects how the world appears, and how the world appears affects how we think of ourselves.[7] But these dimensions can come apart in everyday self-ignorance, more motivated self-deception, and aspiration. When an aspirant like Samantha changes how she thinks about herself and commits to change, she creates a tension between her reflective practical identity and her phenomenological practical identity in hopes of changing how she sees and lives in the world. She hopes that, in time, the change will become more complete and less effortful. But this change is often difficult.

One reason it is difficult is that aspirants can continue to see the world (partly) through the lenses of identities they are trying to leave behind, which conflict with their new self-conceptions—for example, continuing to see tense

student who takes a required course, she argues that the student does act autonomously because his practical identity is a student (105–7). In *Self-Constitution*, she adds, "We have many particular practical identities and so we also face the task of uniting them into a coherent whole" (21). In *Self-Constitution*, she sometimes discusses the [whole ...] in terms of personal identity, but she also sometimes uses these terms interchangeably: "We are each faced with the task of constructing a peculiar, individual kind of identity—*personal or practical identity*" (19–20, emphasis added).

[6] Although I describe particulars of Korsgaard's and Crowell's accounts, I am not arguing for the particulars of either account. One could flesh out the details of these dimensions differently while retaining the core insight of the multidimensional picture.

[7] Thanks to an anonymous referee for helping me think through these relationships in terms of coordination.

situations as calling for them to apologize. When they find themselves with these *residual practical identities*, they may stop seeking to change and instead turn to something else, such as resignation, denial of responsibility, self-deception, or some combination of these.[8] Some former aspirants come to identify fully with their aspirational reflective practical identity and deny responsibility for actions that do not fit with that self-conception. They might see those actions as not really theirs, even though the actions fit with their lagging phenomenological practical identity. Others resign themselves to their situations as if they have no say in (and thus no responsibility for) the matter. Of course, former aspirants do not have a monopoly on resignation, denial, or self-deception. But the same features that make aspiration and responsibility-taking disavowals possible also provide the scaffolding for responsibility-avoidant (pseudo-)disavowals, so we may confuse aspiration with denial or self-deception.[9] We could see the backsliding aspirant as merely being in denial or see the person in denial as an aspirant not yet living up to their aspirations. But lumping these together conceals important differences in these agents.

It is natural to think of who we *really* are either in terms of our aspirations (or core values) or in terms of our actions. But without proper nuance, these both distort our moral psychology. As beings who care about who we are, we are (sometimes thankfully) more than our actions. But we are also (sometimes frustratingly) more than our aspirational selves. The natures of human selfhood and agency depend on our ability to care about and take responsibility for ourselves, including parts of our selves over which we do not have complete immediate control. In this paper, I propose a multidimensional account of practical identity and use Merleau-Ponty's account of world polarization to explore the dynamic between reflective practical identity and phenomenological practical identity. This conceptual framework illuminates the unique profiles of self-ignorance, resignation, aspiration, and denial. It also explains a form of practical ambivalence common in aspirants transitioning from one way of being to another.

1. WHO ARE YOU, ANYWAY?

On the one hand, when Samantha commits to change by asserting, "I'm not going to be that person anymore," she changes who she is. She now values herself under a different description, or at least disvalues herself under some description. But in aspiring to not be "that person anymore," she wants more

[8] This is not to say that all aspirants should continue on their trajectories. They could determine, for example, that their initial aspirations were naïve, misguided, or not worth the costs.

[9] Thanks to an anonymous referee for helping articulate this relationship.

than to see herself differently. She wants to change how she sees and lives in the world. She wants to become someone who does not feel the need to apologize (and does not apologize) when she is not wrong. But she has probably already changed. She has likely become someone who would be disappointed if she were to apologize when not wrong. To make sense of the complexity of Samantha's situation—the change she has wrought and the further change she seeks—let us start with Korsgaard's and Crowell's accounts of practical identity.

1.1. Reflective Practical Identity

As Korsgaard describes the human condition, we find ourselves with impulses—feelings, beliefs, and desires that impel us to act.[10] The reflective structure of human consciousness, however, allows us to control which impulses lead to action. This structure makes autonomy possible, but it brings with it a kind of necessity. It makes it so we *can,* but also *must,* decide which impulses we will endorse (and act on) and which we will reject.[11] And this is where practical identities are crucial. Practical identities—descriptions under which we value ourselves—provide criteria for determining what counts as a reason, for distinguishing impulses we approve of from those we do not: "We endorse or reject our impulses by determining whether they are consistent with the ways in which we identify ourselves."[12]

Korsgaard sees endorsing reasons for action as an act of existential significance. Practical identities are largely socially received roles and ways to live:

> Some we are born into, like being someone's child or neighbor or being the citizen of a certain country. Some we adopt for reasons, like joining a profession that is worthwhile and suits your talents or devoting yourself to a cause in which you ardently believe. Many we adopt voluntarily, but without reasons in anything more than a minimal sense.[13]

[10] I highlight Korsgaard's account because Crowell frames his Heideggerian account as a contrast to Korsgaard's, but also, more substantively, because hers is a paradigmatic account of reflective practical identity, in which self-conceptions play a decisive role in determining one's agential standpoint. One could change details about the structure of the self, the nature of self-conceptions and how they factor into agency, and so forth, while still having reflective practical identity (broadly construed) play a crucial role.

[11] Korsgaard, *Sources of Normativity*, 113.

[12] Korsgaard, *Sources of Normativity*, 120. Sometimes Korsgaard describes endorsement in self-consciously reflective ways, as when she equates endorsement with an agent identifying with the reasons and obligations relevant to some description. Other times, however, she describes it more pragmatically: we endorse some desire, role, or identity when we act in accordance with it. See, for example, *Self-Constitution*, 43.

[13] Korsgaard, *Self-Constitution*, 23.

When we endorse reasons stemming from an identity, we "carv[e] out a personal identity for which we are responsible"—something that Korsgaard sees as "one of the inescapable tasks of human life."[14] When we see ourselves as parents, lovers, teachers, students, or friends, these roles become reason-giving.[15] We commit to (try to) act according to the norms of that identity. When I then act according to impulses that are consistent with those norms, I am likely to "regard [the] movement . . . as *my action*" and not merely as "some force that is at work *on* me or *in* me."[16] When someone who values herself as a student raises her arm to make a comment, she is behind the wheel differently than if her arm rises because of a muscle spasm—*she* does it, one might say, and her reflective practical identity as a student gives her the reasons to do so.[17] For Korsgaard, "autonomy is commanding yourself to do what you think it would be a good idea to do, [which] depends on who you think you are."[18] I get something right, agentially speaking, when I act on reasons flowing from an identity with which I identify. Because Samantha identifies as someone who does not back down, when she rejects the impulse to apologize and instead acts on the impulse to stand her ground, the movement appears to her as an (autonomous) action and not merely the result of forces working on or in her.

However, siding with an impulse does not ensure that it leads to action. We can be moved by impulses that conflict with our adopted practical identities. Addiction and depression, for example, can compromise agency by making some impulses effective even when we do not see them as good reasons and other impulses ineffective even when we see them as good reasons. Samantha's case is different, however. If she apologizes when she is not wrong, the apologetic impulses are consistent with an identity that once more fully (perhaps with her approval) shaped her outlook but that she is now trying to leave

[14] Korsgaard, *Self-Constitution*, 24.

[15] On his reading of Korsgaard, Velleman identifies self-conception with practical identity: "Willing the law is a matter of adopting a self-conception, or 'practical identity'" ("Willing the Law," 297).

[16] Korsgaard, *Self-Constitution*, 18 (emphasis in original).

[17] Frankfurt uses the example of someone raising her arm and an arm rising spasmodically to distinguish between actions and physical movements that happen to her ("Identification and Externality," 58). Kieran Setiya contests this account, at least in regard to reflexes. He contends that my arm moving as a reflex is not something that happens to me but is something that I do as a reflex action. The contrast—something that I do not do—is someone lifting my arm, perhaps during a medical examination. According to Setiya, the ways Korsgaard and Frankfurt distinguish between things I do and things that happen to me already assume we are looking for a certain kind of action—actions done for reasons—and not for what makes something an action (*Practical Knowledge*).

[18] Korsgaard, *Self-Constitution*, 107.

behind. After her explicit disavowal, her practical identity involves a complexity different from the "jumble of conceptions" Korsgaard mentions.[19] Samantha has neither fully left behind who she once was nor fully become who she wants to be. Her reflective practical identity *is* (partly) someone who does not apologize when she is not wrong. But when she backslides, she seems *also* to still be someone who does apologize when she is not wrong. If her identity were to depend entirely on how she explicitly values herself, however, her backsliding apologies would seem to be mere movements, not actions. But given her disavowed identity's influence on how she views the world, seeing her apologies as mere movements seems to mischaracterize both those apologies and her agential situation.

1.2. Phenomenological Practical Identity

Crowell presents his Heideggerian account of practical identity as an alternative to Korsgaard's.[20] He argues that Korsgaard's account relies on an overly reflective and rationalistic picture of human agency, which ultimately leads to a "rationalistic distortion in her phenomenology of action," particularly regarding the everyday coping that characterizes much of our lives.[21] Everyday coping, in Hubert Dreyfus's interpretation of Heidegger, refers to the way we skillfully yet non-deliberatively respond to situations—when driving, for example.[22] Because reflectiveness is often deemed a (sometimes *the*) distinctive feature of humans, Crowell's criticism and his Heideggerian proposal appeal to the intuition that many ordinary actions seem unreflective.[23] Once we possess the relevant skills, many actions rarely seem reflective or deliberative. Without much thought, we respond to what the situation calls for—we loop shoelaces, tap the blinker down to signal a turn, or offer a helping hand or comforting word. Crowell thinks an account of practical identity based on Heidegger's notion of *Worumwillen* (for-the-sake-of-which) can better account for how practical identities function in everyday coping.[24]

[19] Korsgaard, *Sources of Normativity*, 101.

[20] Crowell, "The Existential Sources of Normativity."

[21] Crowell, "The Existential Sources of Normativity," 241.

[22] See Dreyfus, *Being-in-the-World* and *Skillful Coping*. For a clear summary of Dreyfus's account of everyday coping, see Mark Wrathall's introduction to *Skillful Coping*.

[23] Michael Bratman, for example, describes reflectiveness as central to human agency (*Structures of Agency*).

[24] Crowell, "Existential Sources of Normativity." When necessary to disambiguate, I refer to Korsgaard's notion as *reflective practical identity* and Heideggerian *Worumwillen* as *phenomenological practical identity*. Otherwise, *practical identity* refers to both.

For Heidegger, a *Worumwillen* (or for-the-sake-of-which) is “a self-interpretation that informs and orders all my activities.”[25] It is “a possible way of being a self” (such as being a parent, teacher, or carpenter) that organizes or grounds intentions and actions by providing criteria according to which some actions are self-determined, autonomous, expressive of what is my own and not mere happenings in my life.[26] When, as a carpenter, I hammer in nails to secure boards, the “in-order-to” (*Um-zu*) of securing boards is grounded in a “for-the-sake-of-which” (being a carpenter) that I have seized upon.[27] On Crowell’s account, when I try to be some practical identity, it affects how the world appears to me: “When I try to exercise the skills that define [a particular *Worumwillen*], try to live up to the demands of the job, I act for the sake of a possibility of my own being, and only so can things present themselves to me in light of *their* possibilities.”[28] As I try to engage with the world as a carpenter—to live up to the constitutive norms of being a carpenter—the wood, nails, saws, and planes show their distinctive possibilities. The world appears to me as it does to a carpenter.[29]

Maurice Merleau-Ponty describes this phenomenon as world polarization. Under normal circumstances, he writes, a “person’s projects polarize the world, causing a thousand signs to appear there, as if by magic, that guide action, as signs in a museum guide the visitor.”[30] Our projects affect the salience and

[25] Dreyfus, *Being-in-the-World*, 95.

[26] Crowell, “Existential Sources of Normativity,” 242–44. This sounds a lot like Korsgaard, because even though I use Crowell to highlight a different dimension of practical identity, Crowell frames his account as a corrective to Korsgaard (“Existential Sources of Normativity,” 241).

[27] Crowell, “Existential Sources of Normativity,” 244. See Heidegger, *Sein und Zeit*, 86 (Macquarrie and Robinson, 119).

[28] Crowell, “Existential Sources of Normativity,” 245. Of course, one can hammer nails to secure boards “for-the-sake-of” various practical identities—I can hammer nails to help a friend, for example. Hammering nails would then have the same “in-order-to” but be anchored in a different “for-the-sake-of-which.” Thanks to Mark Wrathall for suggesting this possibility.

[29] On Korsgaard’s view, identifying with some role (hopefully) changes how relevant impulses appear; on Crowell’s view, trying to inhabit some role (hopefully) changes how the world appears.

[30] Merleau-Ponty, *Phenomenology of Perception*, 115. Stephan Käufer (“Heidegger on Existentiality, Constancy, and the Self”) and Mark Wrathall (“Who Is the Self of Everyday Existence?”) both draw on Heidegger to develop accounts of the self in terms of polarization (or something very similar). Käufer describes the self in terms of the ability to let ourselves be drawn in by what beckons us (466). Wrathall describes the self as “a function that needs to be performed if a situation is to invite and sustain action: I am the polarization of the affordances of a situation into particular solicitations to act (22). On key points, both Käufer and Wrathall use Merleau-Ponty’s analysis to develop or clarify their Heideggerian accounts of selfhood.

affective orientations of potential actions. Some opportunities are strongly inviting. Others are weakly repulsive. And some fail to stand out. As a parent and professor, helping my child with homework, revising papers, and preparing lectures are more salient than practicing Beethoven's "Moonlight Sonata" (assuming I am not also a pianist with an upcoming recital). Combining Crowell's Heidegger and Merleau-Ponty, whether and the extent to which we have some (phenomenological) practical identity depends on the world being polarized by that identity, with situations "calling for or ruling out certain actions."[31]

World polarization can be sudden and far-reaching. Some new parents quickly see the world in comprehensively parental ways, with the world strongly soliciting actions not previously on their radars. But world polarization is not always that way. Other parents change their explicit priorities but struggle to develop the skills and dispositions that allow them to see and respond well to parenting situations. Our self-conceptions and conscious commitments to projects typically affect how the world appears to us—seeing myself as someone's friend, for example, likely shapes how their struggles or successes affect me—so there is usually coordination between reflective practical identity and phenomenological practical identity. But there can be slippage between the two.

A key difference between Korsgaard's and Crowell's accounts is that, for Crowell, our original self-awareness of our practical identities comes through what we do and how the world appears to us. "I am constantly self-aware," he writes, "because I discover myself in what I do: I am aware of myself as a carpenter, father, or teacher because the things that surround me show me the face that they show to one who acts as a carpenter, father, or teacher does."[32] With a nod to Korsgaard, Crowell elaborates, "to understand oneself as a carpenter, philosopher, father, or friend is not to represent oneself under a certain description but to be able to be those things."[33] Self-understanding, in this sense, is to have skills and dispositions that allow one to effectively navigate the world of an identity. Merely conforming to the relevant standards is insufficient, however. I must *try* to be a carpenter, philosopher, or parent. And trying cannot be reduced to "the exercise of any set of practical skills or abilities" but "presupposes the possibility of acting in light of norms and not acting merely in conformity with norms."[34] Having some practical identity depends on being able to

[31] Rouse, "Self-Awareness and Self-Understanding," 166.

[32] Crowell, "Existential Sources of Normativity," 247.

[33] Crowell, "Competence over Being as Existing," 81.

[34] Crowell, "Competence over Being as Existing," 82. Crowell uses the example of taking a picture with an old camera to illustrate the ability to act in light of norms. Whether the camera is appropriate depends on my purposes: "If I experience the camera as defective because my photographs are blurry, this is because I am trying to capture the moment for

act in light of the difference, relative to that identity, between better and worse, success and failure, and exercise the relevant "ability-to-be" (*Seinkonnen*).[35]

Recall that for the dimension of practical identity that Korsgaard highlights, how we identify is crucial: by adopting and identifying with roles, their norms become normative for us. One could think of the difference between reflective and phenomenological practical identities as highlighting the distinction between (a) seeing and understanding reasons to act and (b) being solicited to act. If I identify as a parent but do not readily see the world as a parent, I still have reasons to act as a parent (in light of my adopted identity) and can act deliberately according to the norms of that identity. But I need to act deliberately because the world does not (yet) solicit me as it would if I more fully had the identity of parent.[36] Conversely, as we see in backsliding aspirants, the world can solicit me to act according to an identity whose deliberative force I have disavowed.

1.3. A Multidimensional Account of Practical Identity

I propose that if we take Korsgaard's and Crowell's (Heideggerian) accounts not as competing but as highlighting distinct but interwoven aspects of practical identity, the resulting multidimensional account of practical identity allows us to better see the distinct contours of self-ignorance, aspiration, resignation, and denial. Moreover, such an account allows us to accept Crowell's broad Heideggerian point that "the greater part of our practical lives" is spent in pre-reflective, non-deliberative action and that primary self-awareness comes not through reflection or introspection but through how the world appears to us, while still holding that the capacity to reflect (for example, on who we are and want to be) still shapes our experience, for good and ill.[37] Thoughts about our values and aims can make certain possible actions more salient to us and others less salient.[38] But those same capacities also allow reflective self-awareness to pull apart from phenomenological self-awareness in ways that underpin self-ignorance, self-deception, denial, and bad faith.

Because the practical outlooks of aspirants are often (partly) shaped by residual practical identities, it can take time, work, and often luck for commitments like Samantha's to take hold, for the influence of disowned identities to

my family. If I experience the same camera as quite suitable, this is because I am trying to make the prototype for a painting in the style of Gerhard Richter" ("Responsibility, Autonomy, and Affectivity," 216).

[35] Crowell, "Responsibility, Autonomy, and Affectivity," 226.

[36] Thanks to Mark Wrathall for highlighting this point for me.

[37] Crowell, "Existential Sources of Normativity," 257.

[38] Komarine Romdenh-Romluc uses the work of Merleau-Ponty to describe different roles that thought can play in action ("Thought in Action").

dissipate and for aspired-to identities to permeate or more fully polarize one's world. As a result, when people say, "That's not who I am," but their actions suggest otherwise, it can be hard to tell backsliding aspirants apart from those who are self-ignorant or in denial. These individuals all have tensions between their behavior (or actions) and their beliefs (or statements of commitment or disavowal), but their situations are different. As aspirants seek to change how they pre-reflectively live in the world, they can experience a sort of practical ambivalence because the polarizations of their world are in flux. But as aspirants, they acknowledge and are working through that ambivalence. The others, by contrast, either are not in positions to enact such a change or choose not to do so. To see how the multidimensional account of practical identity makes sense of a range of cases, we will discuss Samantha, Shakespeare's *Othello*, and several variations of a parent with workaholic tendencies.

1.4. Self-Ignorance and Practical Identities

When there is a gap between how someone sees themselves and how we (and others) see them, it is easy to chalk it up to willful (or semi-willful) self-deception. But active convincing need not be involved. Let us take Eric Schwitzgebel's claim that "[we] live in cocoons of ignorance, especially where our self-conception is at stake."[39] In many cases, the cocoons of ignorance related to self-conceptions (whether called self-ignorance or self-deception) boil down to a mismatch between the self-awareness of phenomenological practical identity—in which one pre-reflectively senses how to respond to different situations and can act accordingly—and the self-awareness of reflective practical identity.[40] We can notice and respond to the world's solicitations without seeing ourselves as having the identities that are tied up with that solicitational structure.

The specific contours of our first-personal experience depend on our roles and characteristics, but our original experience of those roles and characteristics is distinct from (and prior to) our conscious thoughts about whether we inhabit some role or possess some characteristic. Describing how we pre-reflectively experience our identities, Merleau-Ponty writes:

> I am for myself neither "jealous," nor "curious," nor "hunchbacked," nor "a civil servant." We are often amazed that the disabled person or the person suffering from a disease can bear the situation. But in their own eyes they are not disabled or dying. Until the moment he slips into a

[39] Schwitzgebel, "Self-Ignorance," 197.

[40] Herbert Fingarette similarly argues that we can notice things and guide our behavior accordingly without explicitly focusing our attention on them ("Self-Deception Needs No Explaining").

> coma, the dying person is inhabited by a consciousness; he is everything that he sees.... [Our particular characteristics] are the price we pay, without even thinking about it, for being in the world.[41]

His claim that "I am for myself neither 'jealous,' nor 'curious,' nor 'hunchbacked,' nor 'a civil servant'" does not mean that he—or, by extension, each of us—is none of those things. The claim is that even if I assent to a proposition that I am (an) *X*, I still cannot be, for myself, (an) *X*. It is not just that our thoughts about ourselves in terms of such descriptions are phenomenologically secondary. The stronger claim is that we cannot relate to ourselves as (mere) objects to which those qualities, characteristics, or identities apply.[42] Our characteristics and identities shape the contours of our being-in-the-world—polarizing our worlds, for example—but our fundamental first-personal experience of our being-in-the-world is of the specific world made possible by those identities.

In this process, our being-in-the-world typically affects our judgments, including judgments about ourselves. Merleau-Ponty writes: "'I exist as a worker' or 'I exist as a bourgeois' first, and this mode of communication with the world and society motivates both my revolutionary or conservative projects and my explicit judgments ('I am a worker,' or 'I am a bourgeois')."[43] Because our ways of being and styles of life motivate our explicit judgments—including judgments about the sorts of people we are—judgments about ourselves tend to track our being-in-the-world. However, those judgments are secondary and can be clouded by various factors.[44] Simine Vazire suggests that our ability even to see ourselves accurately is worse when the self-perception is of traits thought to be highly desirable or undesirable, such as intelligence and creativity.[45] When we reflect, introspect, or think about what we are like, we can be

[41] Merleau-Ponty, *Phenomenology of Perception*, 458–59.

[42] Thanks to Mark Wrathall for helping to develop this point.

[43] Merleau-Ponty, *Phenomenology of Perception*, 469.

[44] There are various explanations for self-ignorance about practical identities. Lear describes unconscious practical identities (*A Case for Irony*); John Doris claims that we "contain unaccessed and unruly depths" (*Talking to Ourselves*). Some think we have good reason to quiet beliefs that go against our self-conceptions. V. S. Ramachandran suggests that certain kinds of self-deception and confabulation have evolved as tools for imposing "stability, internal consistency and coherence on behavior" (*Phantoms in the Brain*, 254). To maintain stability, we can use "the so-called Freudian defenses—the denials, repressions, and self-delusion that govern our daily lives" (Ramachandran, 134). Frankfurt describes self-deception as an (ultimately misguided) attempt to escape the volitional ambivalence that can threaten robust human agency ("The Faintest Passion"). When our wills are divided, we may avoid volitional stalemate by telling ourselves that we do not *really* care about one of the competing desires.

[45] Vazire, "Who Knows What about a Person?"

(and perhaps inevitably are) imperfectly aware of the practical identities that polarize our worlds.

Let us look at a couple of cases of potential self-ignorance of (portions of) one's phenomenological practical identity. Shakespeare's *Othello* offers a famous illustration of Merleau-Ponty's claim that the jealous person is not, for himself, jealous. If jealousy is a "green-eyed monster," the jealous lover could see a green-tinted world—suspecting his lover of infidelity—while thoroughly unaware of the tint and without thinking of himself as jealous.[46] He could have the phenomenological practical identity of a jealous lover without seeing himself as one. Of course, he *could* see himself as a jealous lover. But his thinking of himself as jealous is secondary to his inhabiting a world characterized by pervasive suspicion, wanting what others have, or dissatisfaction with his situation.

Next, imagine someone who says yes to a request to stay late at work the same night as a family event he has promised to attend. In the first version of the scenario, the parent is a *self-ignorant workaholic* who is reflectively unaware of the global way in which work polarizes his world. Even though he believes he has struck a balance between different areas of life, he sees the world primarily and pervasively in work-oriented ways. He seeks and takes on extra tasks, stays needlessly late at the office, overly diligently checks his email, and so forth. His actions respond so exclusively to work-related solicitations that it creates tension with his reflective practical identity (as someone with good work-life balance). This tension between his explicit self-conception and the way the world is polarized puts him in a position of self-ignorance. Although this tension sometimes leads to some cognitive dissonance, an intervention or a crisis may be needed for him to become reflectively aware of how his world is polarized and of the nature of the tension.

When such agents inhabit worlds polarized by identities with which they do not identify, the miscoordination between the different dimensions of practical identity is often described as self-ignorance or self-deception.

1.5. *Out of the Garden: Responding to Lost Ignorance*

These "cocoons of [self-]ignorance" resulting from miscoordination between explicit self-conceptions and world polarization can be threatened (often helpfully) in various ways. Sometimes we do it largely on our own, through reflection or introspection. But often we need others to point it out to us or to help us talk through them. However it happens, though, once we become better aware of both dimensions of our identity—including undesired characteristics

[46] Shakespeare, *Othello*, 3.3.170.

or phenomenological practical identities—pure self-ignorance is off the table. We then have various options, including:

1. Resignation: We can resign ourselves to the unsavory identity ("*That's just who I am*").
2. Aspiration: We can become aspirants and seek to change so that there is better harmony between our self-conceptions and the way our worlds are polarized ("*That's not who I want to be*" or the aspirational version of "*That's not who I am*").
3. Denial: We can deny that we are that sort of person or that there is any tension between these dimensions (the denial-laced "*That's not who I am*").

And we can and often do cycle between these and other responses.[47]

To illustrate these possibilities, let us look at three variations of our workaholic.

> *The resigned workaholic* is reflectively aware of the global way in which work polarizes his world. He knows he works more than he would like, given other things he cares about. He is sometimes disappointed with himself for failing to do what he thinks is most important. But he has come to accept that he just is someone who works too much. He is not sure if it is ambition, insecurity, an unassailable work ethic, or a combination of these, but he has resigned himself to the situation. Even though he sometimes wishes he could do things differently, he thinks it is beyond his power to enact meaningful change, so he does not try.

In another variation, the individual already aspires to better work-life balance.

> *The aspirant* is a self-acknowledged recovering workaholic. Having recognized his tendency to say yes to projects and to work in ways that conflict with his broader values, he has committed to strike a better balance between work, family, and other pursuits. And he is working to change. The request to stay late finds him inhabiting a world whose polarization is in flux—transitioning away from one in which work is always most salient and toward one in which other interests sometimes take priority. The world still solicits staying late, and he feels the motivational pull, but it now also solicits him to be with his family. When he says yes and stays at work, he is conflicted. Staying late is intelligible to him—it responds to how the world has been (and to some extent still is) polarized. But the

[47] One could, for example, embrace the (previously) undesired characteristic or identity, perhaps revising one's beliefs about its desirability.

> world's polarization is now less straightforward. Nonwork projects are more salient than they once were, even if his aspired-to practical identity and its values, cares, and dispositions are not yet as fully integrated as he would like.

Instead of aspiration or resignation, one could also turn to denial.

> *The workaholic who is in denial* on some level realizes that he has workaholic tendencies and that these tendencies conflict with his other values. His friends and family regularly and (to him) irritatingly call him a workaholic. But rather than considering the possibility and, potentially, working to change, he denies the tendencies and insists to himself and others that he already has balance. As in the case of the self-ignorant, the tension between his phenomenological practical identity and his reflective practical identity persists. But he is more aware of the tension than is the self-ignorant. As in the case of the aspirant, he senses something is off-kilter, and his explicit thoughts and self-presentations have a performative element. However, unlike the aspirant, his thoughts and denial aim not to change but to perpetuate his current way of being. But there is also some resignation. Rather than seeking to change, he rejects the comments of friends and family and muffles his hunch that they might be right. If he continues in his denial, his reflective practical identity could become increasingly ineffectual, insensitive to the way the world appears to him.

Even though the person stays at the office in all variations, these individuals are in different agential situations. They do have some things in common. For example, if asked about being a workaholic, the self-ignorant, the aspirant, and the denier might all claim, "That's not who I am." And apart from the self-ignorant individual, they are all aware of miscoordination between the different dimensions of their identities. But even the more self-aware individuals respond to that self-awareness differently. When the aspirant and the denier say, "That's not who I am," one says it to reinforce a commitment to change and the other to avoid change. The resigned individual and the person in denial, then, each fails to properly account for a key aspect of human existence. And even though they fail to account for different aspects, they are both in bad faith. They both fail to take responsibility for the selves that they are—one by denying that their phenomenological practical identity is genuinely them, the other by denying that they (likely partly through their reflective practical identity) have the capacity to shape how the world appears to them.[48]

[48] In Sartre's famous account, bad faith is possible because human beings have the twofold property of being both "a facticity and a transcendence" (*Being and Nothingness,*

Some degree of self-awareness—including an awareness of the different dimensions of one's practical identity and the level of coordination between them—is typically a precondition for taking responsibility for who we are. Such self-awareness can facilitate self-governance and promote coherence among our values.[49] As we are better aware of how the world solicits us and how we (tend to) respond to those solicitations, we may be better able to respond to situations and, over time, to shift how we are solicited. This could ultimately help us better coordinate our different identities (such as parent and writer) and the different dimensions of those identities. But of course, improved self-awareness does not always lead us to take responsibility for ourselves. All too often, improved self-awareness leads to denial and the avoidance of responsibility. Taking responsibility can be difficult, perhaps partly because it can be difficult to change how the world is polarized for us.

1.6. World Polarization, Residual Practical Identities, and Practical Ambivalence

When we aspire to become something, we want not only to act a certain way but to bring about a "deep change in how one sees and feels and thinks."[50] We want our practical outlook and actions to naturally and seamlessly reflect the new identity. If I aspire to be a good parent, I want to see, feel, and think in a way that allows parental actions to be largely non-deliberative or "spontaneous," to borrow Brownstein's term.[51] Typically, we can respond skillfully yet non-deliberatively to situations only when the relevant norms and dispositions are so thoroughly incorporated into our way of being that they have become muscle memory, so to speak, and no longer require reflective deliberation.[52] The agent

99). According to Sartre, "These two aspects of human-reality are in truth—and ought to be—capable of being validly coordinated. But bad faith does not want to coordinate them" (99). Bad faith involves what Wrathall calls "a motivated failure to see" that we are responsible for the disintegration or lack of coordination between facticity and transcendence ("Ambiguity, Opacity, and Sartrean Bad Faith," 287). Sartre's accounts of human existence and, by extension, bad faith are complex, and I do not argue here that the miscoordination described in my multidimensional account is identical to the miscoordination between facticity and transcendence that Sartre describes. I am suggesting, however, that there are (at least) relevant structural parallels, and that the multidimensional account of practical identity may help illuminate different varieties of bad faith. For more on Sartre's account of bad faith, see *Being and Nothingness*.

[49] Or at least help us be more clear-eyed about the tensions in our lives.

[50] Callard, *Aspiration*, 2.

[51] Brownstein, *The Implicit Mind*.

[52] Driving and typing are common examples of skills that start with very deliberate learning but, once one becomes skilled, can appear as very nuanced "autopilot." To be clear, practical identities need not begin with reflective deliberation. Some are largely the result of

who has "become" the identity now sees the world in light of the identity and can skillfully respond without deliberating about what norms apply in a situation. To become (or be) a certain kind of person, then, is to have (or develop) a certain practical outlook and to be able to inhabit the world accordingly.

The process of aspiration—of becoming (and then being) a certain kind of person—is thus often fraught and ongoing because successful aspiration involves changing not only how we see ourselves but also how the world is polarized for us. To effect these changes often involves not only acquiring new identities but also turning away or distancing ourselves from identities that, to that point, have shaped how we saw, felt, and thought about the world. Attempts at distancing can take many forms, but they sometimes involve explicitly disavowing previous identities or ways of being, as in Samantha's "I'm not going to be that person anymore." Merely verbal disavowal is obviously inadequate and can be worse, potentially involving deception of others or ourselves (and perhaps both) in the service of perpetuating bad behavior. But even genuine, clear-eyed attempts at disavowals can be frustratingly ineffectual, failing to eradicate the influence of renounced identities. The uptake of aspired-to identities can be slower and more difficult than we anticipate. Despite our best efforts, renounced identities can continue to affect, even hold sway over, our practical outlook. In short, we often find ourselves with *residual practical identities,* identities that continue to polarize our worlds even though we no longer value ourselves under the relevant descriptions.

Take Gary Watson's example of someone who "thinks his sexual inclinations are the work of the devil, that the very fact that he has sexual inclinations bespeaks his corrupt nature."[53] Even if this person were to stop believing that his sexual inclinations are the work of the devil or signs of his corrupt nature, his world could continue to be polarized in a way that makes sexual actions appear repulsive. The world would then solicit him according to an identity he no longer identifies with, and which could be at odds with his other reflective and/or phenomenological identities.[54] Or, turning to our earlier examples, even if Samantha no longer believes that she should apologize when she is not wrong, her world could still be polarized by the identity of one who preemptively and faultlessly apologizes. The world could still solicit her to unfairly take the blame for situations and to apologize to avoid conflict. If she were to

acculturation. And even with actively acquired practical identities, the understanding of relevant norms is often largely inherited from and tacitly shaped by others.

[53] Watson, "Free Agency," 19.

[54] I have described residual phenomenological practical identities, but residual reflective practical identities are also possible. We could, for example, continue to identify with and value ourselves under some description after that identity no longer polarizes our world.

apologize, she would likely be disappointed with herself. But she would be disappointed precisely because the action would make sense to her and would indicate that she has not yet become who she aspires to be. We could say similar things about the person aspiring to better work-life balance who continues to see a thoroughly work-polarized world.

What should we say, then, about those who genuinely disavow some identity and want it to stop guiding their actions but continue to see the world through the lens of and act in keeping with the disavowed identity? In these cases of complex world polarization, the agent's world is polarized by the aspired-to identity *and* the renounced identity. The world could thus simultaneously solicit different actions, or the same action could appear as both attractive and repulsive, leading to a state of practical ambivalence.[55] This complex polarization reflects the complex identities of aspirants. But even practically ambivalent aspirants—whose worlds are polarized in different (perhaps competing) ways—are in a different position than those who are self-ignorant, resigned, or in denial.[56]

Even if aspirants do not or cannot see progress, by identifying with and committing to a different way to live, they have changed their reflective practical identities. And because reflectively endorsing or subjectively identifying with an identity often changes how our worlds are polarized—by changing our attention and changing the world's solicitations, as well as the strength of those solicitations—changing one's reflective practical identity already tends to change one's phenomenological practical identity. But until a more complete repolarization takes hold for these work-in-progress aspirants, these changes are usually partial and effortful, more deliberative and reflective than habitual or second nature. The aspired-to identities have not yet taken hold, and disavowed identities still have influence. Because reflective endorsement and subjective identification are neither necessary nor sufficient for one to robustly have some phenomenological practical identity, effective aspiration often depends on continuing to productively acknowledge and work through residual identities.

I have focused here on the complex world polarization and the resulting practical ambivalence we see in aspirants. But because human agency and

[55] Practical ambivalence can also arise when different reflective identities have competing demands. For example, consciously identifying as both parent and professor could lead to competing pressures. But even if there are potential pressures between two (or more) competing reflective identities, practical ambivalence in aspirants is distinct because one experiences practical ambivalence even though one's identity is settled on the reflective level.

[56] Another possibility is a type of radical acceptance of the sort of person one is. Like resignation, acceptance involves a recognition of one's phenomenological practical identity. But it does not completely cede control over the situation and could be preliminary to aspiration.

selfhood involves negotiating and coordinating the different dimensions of ourselves, these phenomena are common and not isolated to aspirants (at least insofar as Agnes Callard understands aspiration). Even maintaining some identity—say, being a parent or teacher—is an ongoing process in which we interpret the identity and its contours, inhabit a world polarized by that identity, learn from others, and, as we live in the world, revise our understandings of our values and identities, and on and on. For these reasons, options like denial and resignation always lurk around the corner. When we find ourselves doing things that we know we should not or wish we would not, rather than facing the tensions between the different dimensions of our practical identity and working to better coordinate them, it is all too easy (or all-too-human) to preserve the tension or gap by denying either its existence or our capacity to close it.

2. Aspiration, Denial, and Resignation

2.1. That's Not Who I Am

Just as Samantha says, "I'm not going to be that person anymore," we can say, "That's not who I am. From this moment on . . ." in order to disavow actions reflecting certain values and to reinforce a commitment to a different path or set of values. When we claim to not be "that person anymore" but do what "that person" would do, it can be a way for us, as aspirants, to distance ourselves from our past identities as part of an effort to change our way of being. But we can say the same words while in denial in order to avoid or deny responsibility for our actions. Responsibility-avoidant disavowals simply express denial that some identity applies to us and are decidedly not part of an effort to change.[57] This kind of denial can ultimately impede our ability to become who we want to be or, in some cases, to act as the people we think we already are.

Aspirants and deniers both have a miscoordination between the identities with which they explicitly identify (their reflective practical identities) and the identities that polarize their world (their phenomenological practical identities). As a result, we might call all such agents self-deceived if they were to, like Samantha, claim to "not be that person anymore." But that would be a mistake. The denier's denial that she is a certain way is entwined with a lack of interest in

[57] Specifics matter, of course. Self-ignorant denial is different from denial couched in motivated (somewhat self-aware) self-deception. However, because we can deny only what has been raised as a possibility, if we are in denial, pure self-ignorance is probably off the table. More commonly, as we become better aware of potential tensions between dimensions of our identity—or, more generally, of some unpleasant trait, quality, or tendency—instead of taking responsibility and working to change, we double down on those identities while simultaneously denying that we have them.

changing in order to better coordinate her phenomenological practical identity with her explicit self-conception. By contrast, the aspirant is keenly aware of her phenomenological practical identity and is trying to change herself in order to better coordinate how she sees the world with how she (aspirationally) thinks of herself. When she says, "That's not who I am," she is not attempting accurate self-assessment so much as stating a changed self-conception and committing to be (or not be) a certain kind of person. She is not avoiding the difficult path of changing her orientation toward the world but reinforcing her commitment to that change.[58]

This is why aspirant Samantha, as an aspirant, is different, for example, from someone who claims to be "a nice guy" but who is consistently rude and inconsiderate and shows no effort to change how he treats others. The self-identified "nice guy" could be merely self-ignorant. Or he could be in denial if, when pressed, he refuses to consider his actions or simply reinterprets them to preserve a pleasant self-conception. But Samantha is different. When she apologizes despite her commitment, she acts according to a residual phenomenological practical identity whose influence she is working to leave behind.

2.2. Denial and Resignation

In one form of denial, we identify fully with our reflective practical identities—"That is who I *really* am," for example—and reject that our phenomenological practical identities are genuinely ours. This form of denial simultaneously overvalues and misunderstands the role of reflective practical identity. To be sure, the emphasis on reflective practical identity gets something right. Our capacity for reflectiveness (and, by extension, our reflective practical identity) allows us to shape our phenomenological practical identity; it can help us direct and take responsibility for our ways of being. However, in the form of denial in which we fully identify with our reflective practical identity in a way that detaches it from our phenomenological practical identity, we end up undercutting the influence of our reflective practical identity. When we claim in denial that we are not (or are) that person, we weaken or undermine our ability to be self-responsible with regard to that aspect of our existence. This often appears as self-enhancement, in which we downplay evidence that conflicts with a positive self-conception by exaggerating the good and minimizing the bad. But one could also identify with a negative self-conception and downplay conflicting evidence. Whatever the details, however, insofar as the denier hides from or ignores his phenomenological practical identity, he risks ever-increasing tensions and dissonance between

[58] For recent treatments of aspiration and moral improvement, see Callard, *Aspiration*; and Stohr, *Minding the Gap*.

his explicit self-conceptions and his way of being. Though not always pleasant, acknowledging that unwanted qualities (and residual practical identities) are part of ourselves is often crucial to being able to better work through them. To actively change our way of being requires us not only to be somewhat aware of it but also to take responsibility for it.[59] In denial, we refuse to consider that there could be a miscoordination between our self-conceptions and our way of being. Or if we consider it, we refuse to take responsibility for the dimensions or their coordination. In this way, as Sartre describes it in his discussion of bad faith, "I am in a place where no reproach can reach me."[60]

Although resignation looks different from denial, it is the other side of the same coin. In resignation, the agent is very sensitive to certain features of his phenomenological practical identity but fails to understand or acknowledge that he has some power to bring about a change in how the world appears to him. He resigns himself to the way the world appears to him, as if he were a mere object and not the sort of being who can shape his existence. The resigned workaholic, for example, is keenly aware that work considerations thoroughly polarize his world but fails or refuses to see that how his world is polarized and how he lives in the world depend (at least partly) on how he self-identifies, how he understands his roles, and so forth. In denying (or refusing to see) that how his roles and dispositions shape his practical identity depends partly on how he takes them up, for example, he also denies (or tries to deny) responsibility for himself. He sees his reflective practical identity not as a dimension of himself that can shape and shift his phenomenological practical identity, but as an inert acknowledgment of who he is.[61]

The denier and the resigned individual both separate themselves from the aspirant or the proto-aspirant in the way they avoid responsibility for their whole selves. When someone genuinely asks themselves if they are jealous or a workaholic, for example, they put themselves in a different space than someone who reflexively and emphatically denies being jealous or a workaholic without considering the possibility.[62] Knee-jerk emphatic denial closes

[59] Because of these difficulties, much apparent aspiration could in fact be denial in which we identify with our aspirational self and deny that unwanted features or dimensions of ourselves are our own.

[60] Sartre, *Being and Nothingness*, 100.

[61] Resignation, so described, is distinct from a kind of acceptance that is compatible with self-responsibility. Such acceptance could lead to self-directed aspirational change. But it could also manifest as one embracing and taking responsibility for one's way of being without the deep change usually associated with aspiration.

[62] Take, for example, the question that guides Neil Levy in "Am I Racist? Implicit Bias and the Ascription of Racism." If we understand racial biases as functioning not merely on the individual level but as structuring the social world in various ways, there are unique

us to the possibility of learning about and potentially changing ourselves. The resigned agent's denial of the capacity to change is obvious, but perhaps less obvious is the way in which, like denial, resignation tries to evade responsibility for actions. Although the denier and the resigned individual veer too far in different directions, they similarly (attempt to) avoid responsibility in ways that undermine their agency.

3. CONCLUSION

Earlier we asked, "What sort of person is Samantha?" Specifically, what sort of person is she after she claims "I'm not going to be that person anymore" but then finds herself doing the very things "that person" did? Because a backsliding aspirant can look a lot like someone who is in denial, we might be tempted to say that Samantha is in denial. But we are now better positioned to see why that is not the case. If someone were to catch her apologizing and ask her, "Well, is that the sort of person you are?" before responding, Samantha might first want to ask how much time her interlocutor has. But once that is settled, if we assume that her aspiration and commitment to change is sincere, she can genuinely say, "That's not who I am." At the very least, she has changed her reflective practical identity, an important dimension of who she is. At the same time, a more honest and likely more effective aspirational path would also lead her to acknowledge that the tendency to apologize is also part of who she is, albeit a part that she is trying to change. Whereas the denier refuses to consider the possibility, Samantha fully owns not only that she has been "that sort of person" but that, to some extent, she still is "that person." In an important sense, then, Sartre's Inez is right that what we do shows who we truly are. But to really understand "what we do," we need a nuanced picture of the agent.

Statements of disavowal made by backsliding aspirants are different from those made by those who are in denial. When entangled with denial, they are likely attempts to evade responsibility without genuinely changing or making amends. By contrast, for aspirants, such statements can be a way to take responsibility for oneself or renew commitment to change. Yet, even though reflective practical identities (partly) shape our world and life, explicit aspirations do not automatically take hold and do not exhaust how the world is polarized for us. As Merleau-Ponty puts it, "My freedom, even if it has the power to commit me to [some new cause], does not have the power to turn me immediately into

complexities facing the aspiring anti-racist. Thanks to an anonymous reviewer for highlighting potential differences between different types of aspiration, including the ineliminable social dimension of some things we may aspire to become or leave behind.

what I decide to be."[63] When Samantha says, "I'm not going to be that person anymore," she commits herself to a new path, but that commitment does not yet fully and immediately change her into the person she aspires to be. Until her everyday being-in-the-world more fully reflects that commitment, she is not yet fully one who apologizes only when she is wrong.

For various reasons, deep, multidimensional change is no small feat. For one, the shapes of identities—inherited or actively acquired—are not entirely up to us. To some degree, we are unconsciously socialized into ways of being. Even when we consciously work to inhabit a role and to live a certain way, the dimensions of our identities—the shapes of roles, the way we are disposed to act in different situations, our notion(s) of the good life, and so forth—are largely acquired through upbringing and socialization. The privileged child may not think of himself as privileged. He could attribute his success entirely to his hard work while failing to see how his privilege has served as a boost or safety net along the way. The talented athlete sees specific actions as appropriate responses to situations without realizing how her athletic gifts make actions viable for her that are not for most people. Her pre-reflective experience of herself is not as *exceptionally talented*, but as *take (make) this sho*t or *run past that defender*.[64] Or returning to Samantha, before she realizes that she has been apologizing when she was not wrong and commits to change, her initial experience in relevant situations is of a world calling for apologies in tense situations, not of herself as one who apologizes when she is not wrong. Once aware of the tendency, however, she has options. She can take responsibility for different dimensions of herself and the coordination between them—either embracing the tendency perhaps, or, as we have described her, changing how she thinks about herself in hopes of effecting deeper change.

Frankfurt claims that humans are distinctive because they can want to be different, in their preferences and purposes, from what they are.[65] The multidimensional account goes further. Our aspirations—who and how we want to be—partly constitute who we are. But we are also more than our self-conceptions and aspirations (or reflective practical identities), so it can be an uneasy fit between this dimension of ourselves, on the one hand, and how we see the world and act, on the other. We could be better, worse, or just different. But the nature of those tensions and how we relate to them—particularly, how we seek to resolve (or, in cases of denial, avoid resolving) them—underpin

[63] Merleau-Ponty, *Phenomenology of Perception*, 473.

[64] See, for example, David Foster Wallace's description of Roger Federer in "Federer Both Flesh and Not" (20–21).

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

self-ignorance, denial, resignation, or aspiration. Our self-conceptions and aspirations play crucial roles in human agency. But just as we distort the nature of who we are if we overlook their role in determining the sorts of people we are, there is a parallel risk in overvaluing their importance in the selves that we are.[66]

Brigham Young University
justin_white@byu.edu

REFERENCES

Atkins, Kim, and Catriona Mackenzie, eds. *Practical Identity and Narrative Agency*. New York: Routledge, 2007.

Bratman, Michael E. *Structures of Agency: Essays*. Oxford: Oxford University Press, 2007.

Brownstein, Michael. *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*. Oxford: Oxford University Press, 2018.

Callard, Agnes. *Aspiration: The Agency of Becoming*. Oxford: Oxford University Press, 2018.

Crowell, Steven. "Competence over Being as Existing: The Indispensability of Haugeland's Heidegger." In *Giving a Damn: Essays in Dialogue with John Haugeland*, edited by Zed Adams and Jacob Browning, 73–102. Cambridge, MA: MIT Press, 2017.

———. "The Existential Sources of Normativity." In *Normativity and Phenomenology in Husserl and Heidegger*, 239–60. Cambridge: Cambridge University Press, 2013.

———. "Responsibility, Autonomy, and Affectivity: A Heideggerian Approach." In *Heidegger, Authenticity, and the Self: Themes from Division Two of Being and Time*, edited by Denis McManus, 215–42. New York: Routledge, 2014.

———. "*Sorge* or *Selbstbewußtsein*? Heidegger and Korsgaard on the Sources of Normativity." *European Journal of Philosophy* 15, no. 3 (December 2007):

[66] Earlier versions of these ideas were presented at the Horizons of Phenomenology conference (hosted at UC Merced), the Southwest Seminar in Continental Philosophy (hosted at Northern Arizona), the Phenomenology and Personal Identity conference (hosted at Charles University in Prague), and the Brigham Young University Humanities Colloquium. I am grateful to many at those events for their helpful comments and questions. I am especially grateful to Brynna Gang, Jonathan Pulsipher, Gabbie Schwartz, Niels Turley, Mark Wrathall, and four anonymous referees for their insightful comments and questions on earlier drafts of this paper.

315–33.

Doris, John M. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press, 2015.

Dreyfus, Hubert L. *Being-in-the-World: A Commentary on Heidegger's Being in Time, Division 1*. Cambridge, MA: MIT Press, 1991.

———. *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*. Edited by Mark A. Wrathall. Oxford: Oxford University Press, 2014.

Fingarette, Herbert. "Self-Deception Needs No Explaining." *Philosophical Quarterly* 48, no. 192 (July 1998): 289–301.

Frankfurt, Harry G. "The Faintest Passion." In *Necessity, Volition, and Love*, 95–107. Cambridge: Cambridge University Press, 1999.

———. "Freedom of the Will and the Concept of a Person." In *The Importance of What We Care About: Philosophical Essays*, 11–25. Cambridge: Cambridge University Press, 1988.

———. "Identification and Externality." In *The Importance of What We Care About: Philosophical Essays*, 58–68. Cambridge: Cambridge University Press, 1988.

Heidegger, Martin. *Being and Time*. Translated by John Macquarrie and Edward Robinson. New York: Harper and Row, 1962.

———. *Sein und Zeit*. 1927. 6th ed. Tübingen: Max Neimeyer Verlag, 2006.

Käufer, Stephan. "Heidegger on Existentiality, Constancy, and the Self." *Inquiry* 55, no. 5 (2012): 454–72.

King, Samantha. *Born to Love, Cursed to Feel*. Kansas City, MO: Andrews McMeel Publishing, 2016.

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

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

Lear, Jonathan. *A Case for Irony*. Cambridge, MA: Harvard University Press, 2011.

Levy, Neil. "Am I Racist? Implicit Bias and the Ascription of Racism." *Philosophical Quarterly* 67, no. 268 (July 2017): 534–51.

Merleau-Ponty, Maurice. *Phenomenology of Perception*. Translated by Donald A. Landes. New York: Routledge, 2012

Ramachandran, V. S., and Sandra Blakeslee. *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York: William Morrow, 1998.

Romdenh-Romluc, Komarine. "Thought in Action." In *The Oxford Handbook of Contemporary Phenomenology*, edited by Dan Zahavi, 198–215. Oxford: Oxford University Press, 2013.

Rousse, B. Scot. "Self-Awareness and Self-Understanding." *European Journal of*

Philosophy 27, no. 1 (March 2019): 162–86.

Sartre, Jean-Paul. *Being and Nothingness*. Translated by Sarah Richmond. New York: Washington Square Press, 2018.

———. *No Exit*. In *No Exit and Three Other Plays*. New York: Vintage, 1989.

Schwitzgebel, Eric. "Self-Ignorance." In *Consciousness and the Self*, edited by Jeeloo Liu and John Perry, 184–97. Cambridge: Cambridge University Press, 2012.

Setiya, Kieran. *Practical Knowledge: Selected Essays*. Oxford: Oxford University Press, 2016.

Shakespeare, William. *The Norton Shakespeare*. New York: W. W. Norton, 1997.

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

Stohr, Karen. *Minding the Gap: Moral Ideals and Moral Improvement*. Oxford: Oxford University Press, 2019.

Vazire, Simine. "Who Knows What about a Person? The Self-Other Knowledge Asymmetry (SOKA) Model." *Journal of Personality and Social Psychology* 98, no. 2 (February 2010): 281–300.

Velleman, J. David. "Willing the Law." In *Self to Self: Selected Essays*, 284–311. Cambridge: Cambridge University Press, 2005.

Wallace, David Foster. "Federer Both Flesh and Not." In *Both Flesh and Not*, 5–36. New York: Back Bay Books, 2013.

Wallace, R. Jay. *The View from Here: On Affirmation, Attachment, and the Limits of Regret*. Oxford: Oxford University Press, 2013.

Watson, Gary. "Free Agency." In *Agency and Answerability: Selected Essays*, 13–32. Oxford: Oxford University Press, 2004.

Westlund, Andrea. "Who Do We Think We Are?" *Philosophy and Literature* 43, no. 1 (April 2019): 173-91.

Wrathall, Mark A. "Ambiguity, Opacity, and Sartrean Bad Faith." *Humana Mente* 20 (February 2012): 265–91.

———. "Introduction." In *Skillful Coping*, by Hubert Dreyfus, 1–22. Oxford: Oxford University Press, 2014.

———. "Who Is the Self of Everyday Existence?" In *From Conventionalism to Social Authenticity. Heidegger's Anyone and Contemporary Social Theory*, edited by Hans Bernhard Schmid and Gerhard Thonhauser, 9–28. Cham, Switzerland: Springer, 2016.

Journal of Ethics and Social Philosophy
vol. 26, no. 1 · november 2023

https://doi.org/10.26556/jesp.v26i1.1837

AGNOSTICISM AND PLURALISM ABOUT JUSTICE

Adam Gjesdal

Reasonable citizens and their representatives face a problem of criterial indeterminacy. They have many criteria for deciding how to vote: they can decide on the basis of a coin flip, consult a haruspex, or thoughtfully apply what they justifiably take to be the correct principles of liberal justice. Sometimes it is permissible to use any of these criteria. Yet, it should be fairly obvious that democratic citizens should not treat these criteria as always and entirely on a par. As I will show, many political liberals harbor commitments that prevent them from making this fairly obvious point. Political liberalism treats reasonableness as the core concept for evaluating coercive public policy. Analyses of reasonableness specify a filter through which many implementable public policies pass, generating the problem of criterial indeterminacy for how to rank order the resulting set of feasible, reasonable policies. According to what I call *agnosticism about justice*, any criteria for rank ordering reasonable policies is as good as another. This implies that a haruspex who only reads the entrails of roadkill (hence violating no animal's rights by killing them) is no better and no worse a guide for rank ordering reasonable policies than is John Rawls. Should you find this an absurd result, you should also reject agnosticism. Yet, as I will show, agnosticism is a well-motivated response to reasonable disagreement about justice, and political liberals so far have offered no good alternative to it.

This paper presents that alternative. According to *pluralism about justice*, multiple conceptions of justice are correct, and reasonable citizens should appeal to the criteria of a correct conception of justice when rank ordering reasonable policy.[1] Both agnosticism and pluralism share several features. They abandon an idealized vision of the just society on which all citizens agree and deliberate from the very same conception of justice. Where they differ is that

[1] Pluralism, as I conceive of it here, is not to be confused with the descriptive, sociological claim that liberal societies feature a variety of cultures, associations, jurisdictions, etc. That alternative usage informs Jacob Levy's analysis of liberal orders. See Levy, *Rationalism, Pluralism, and Freedom*, 27. Nor is justice pluralism to be confused with Isaiah Berlin's metaphysical thesis of moral pluralism. See Berlin, "Two Concepts of Liberty," 241–42.

pluralism, but not agnosticism, can answer the question of criterial indeterminacy. On the pluralist analysis, citizens should rank order reasonable policies by appeal to the freestanding, liberal conception of justice they justifiably regard as correct because it is in reflective equilibrium for them. A correct conception of justice, *a*, that is in reflective equilibrium for one reasonable citizen, Alf, need not be in reflective equilibrium for another citizen, Betty. Nevertheless, Alf does his best to treat Betty as a free and equal citizen when he deliberates within *a*'s framework to determine which coercive public policy to support, out of a set of reasonable, feasibly implementable policies.

1. Reasonableness as Filter

Anyone familiar with political liberalism knows that the theory does not traffic in notions of correctness. Justifying a law by appeal to some controversial notion of correctness is tantamount to imposing a view onto citizens they lack sufficient reasons to endorse. Political liberals—unlike perfectionists and comprehensive liberals—see such impositions as illegitimate uses of political power.[2] My aim is to convince you that political liberals should traffic in a notion of correctness—which, as I will show, is related to but distinct from a notion of truth—and that they can do so without collapsing into either perfectionism or comprehensive liberalism. First, we need to get in view political liberalism's core concept of reasonableness. This section follows Jonathan Quong's analysis, seeing reasonableness as a filter on arguments justifying coercive public policy. Many policies can be reasonable, leading to the problem of selecting a single reasonable policy to support, which we turn to in the next section.

Political liberalism treats reasonable disagreement about justice as one of the enduring features of life under free institutions.[3] This is why it treats reasonableness, not justice, as the standard for evaluating coercive law. A law is reasonable when it is publicly justified: when all citizens subject to that law have sufficient reason to endorse it. Public justification can be either a low or a very high bar. The high-bar form sees very few laws as publicly justified.[4] The low-bar form, on which we focus here, sees many laws as reasonable. Jonathan Quong's

[2] Quong, *Liberalism without Perfection*, 3.

[3] Rawls often emphasizes reasonable disagreement about the good; see *Political Liberalism*, xlvii. Authors who place central emphasis on disagreement about justice include D'Agostino, *Free Public Reason*, 23–24; Gaus, *The Order of Public Reason*, 43; Quong, *Liberalism without Perfection*, 6–7; Sen, *The Idea of Justice*, 56–58; Vallier, *Must Politics Be War?* 56–58; Waldron, *Law and Disagreement*, 151.

[4] Gaus's model of public reason exemplifies the high-bar form with its "tilt" toward justifying classical liberal positions (*The Order of Public Reason*, 526).

analysis exemplifies core features of the low-bar form. For Quong, all citizens have sufficient reason to endorse publicly justified law because the reasons supporting that law are "mutually acceptable." Mutual acceptability acts as a filter. Laws justified by appeal to a controversial theory of the good life, or by appeal to religious belief, are filtered out. Laws supported by mutually acceptable justificatory arguments based in freestanding political values are filtered in. The justificatory filter allows many arguments through, without ranking the resulting range of reasonable policies as better or worse.

Political liberals need this filter to distinguish disagreements about justice from disagreements about religion or the nature of the good. Quong sees citizens as sharing a substantial set of political values and principles but disagreeing about how these values and principles should be interpreted, weighed, and balanced to yield substantive policy conclusions. Quong gives us the following example of such a disagreement. Sara and Tony are debating the "(in)justice of allowing the Catholic Church to discriminate on the basis of gender when employing priests."[5] At issue between the two is whether religious liberty should exempt private associations like the church from state interference on employment practices. Although Sara and Tony disagree about the proper interpretation and weight to assign to religious liberty, each recognizes the other as deliberating within a shared framework. Both are appealing to freestanding political values—values that do not presume the truth of any specific controversial moral or religious worldview. As reasonable liberal citizens, both accept these values, and each can recognize the other as employing these values in a mutually acceptable way.

Premises in a justificatory argument are mutually acceptable when they satisfy the following three conditions: "(a) all the parties must be sincere, (b) the conflicting positions must be grounded in free standing political values, and (c) the conflicting arguments must represent a plausible balance of political values."[6] Citizens are sincere when they believe a justificatory argument supports a specific law.[7] Positions are grounded in freestanding values when they are "not presented as derived from, or as part of, any comprehensive doctrine."[8] Quong leaves the standard of plausibility largely unanalyzed. But that standard must allow Sara to regard Tony's justificatory argument as both plausible and mistaken. Both Sara and Tony regard their own balancing of political values as best, or as the "most reasonable interpretation," of these values.[9] Each has reasons for

[5] Quong, *Liberalism without Perfection*, 205–6.

[6] Quong, *Liberalism without Perfection*, 207.

[7] For a more detailed analysis, see Schwartzman, "The Sincerity of Public Reason," 384–87.

[8] Rawls, *Political Liberalism*, xlii.

[9] Quong, *Liberalism without Perfection*, 206.

believing that their favored interpretation and weighting of the value of religious liberty is best. These reasons need not strike the other as convincing. They need only provide "a plausible explanation as to why one public value ought to be prioritized over the other in cases of this kind."[10] On the one hand, this plausibility standard is a virtue of the analysis in that it does not place burdensome justificatory demands on citizens to explain and defend their positions. On the other hand, the plausibility standard leads to counterintuitive implications. It opens Quong's analysis to the problem of criterial indeterminacy, as I argue below.

The mutual acceptability criterion allows many policies to be publicly justified. Let us say that Sara is in support of a law, L_1, that would require the Catholic Church to employ women priests. Tony is in favor of a law, L_2, that exempts the church from discriminatory hiring laws. Assuming a democratic mechanism to be in place for deciding between L_1 and L_2, either of these laws, were they selected by that mechanism, would be publicly justified. This is because the reasons Sara and Tony offer in private discussion supporting L_1 and L_2 show those laws to be mutually acceptable, and therefore reasonable for all citizens. The reasons supporting both laws pass through the "filter" of mutual acceptability, making either law publicly justified—as long as it is selected democratically. Thus, Quong's analysis exemplifies the low-bar form of political liberalism, which sees many laws as publicly justifiable.

On Quong's analysis, political liberals should not expect there to be an agreed-upon public conception of justice. For Rawls, a public conception of justice provides "a shared basis for the justification of political and social institutions" and that articulates and orders society in a "principled way."[11] Instead of sharing a conception of justice, Tony and Sara share reasonableness as a filter on the reasons they use to justify coercive public policy.[12] The shared filter allows society to select either L_1 or L_2, regardless of whether one of these two laws is overall more coherent with the existing body of relevant legislation. Such a society has no "regulative political conception of justice" that either guides citizens' deliberations about justice or provides coherence to the existing body of law.[13] Some political liberals view this as a costly result.[14] Whatever that cost may be, I assume it is worth bearing. In what follows, I assume with Quong that political liberals should permit citizens to act on their private judgments regarding what

[10] Quong, *Liberalism without Perfection*, 209.

[11] Rawls, "The Idea of Overlapping Consensus," 421.

[12] Quong, *Liberalism without Perfection*, 187.

[13] Rawls, "The Idea of Overlapping Consensus," 421.

[14] Gaus and van Schoelandt argue that without public justice the politically liberal society is subject to voting cycles and other forms of incoherence ("Consensus on What?," 170–71).

freestanding liberal justice requires. Such a view must do without public justice and bear the attendant costs of social incoherence. The extent of this social incoherence will depend on two factors: first, how many laws can pass through the mutual acceptability filter, and second, how many of those laws citizens will actually endorse and vote for. This second factor leads us to the problem of selecting criteria for rank ordering mutually acceptable law.

2. AGNOSTICISM AND CRITERIAL INDETERMINACY

How do reasonable citizens decide which of the set of reasonable policies to favor? Quong says that Sara and Tony will favor those policies they regard as "most reasonable."[15] He never offers a substantive analysis of this notion. For the moment, let us say that the most reasonable policy is the one justified by appeal to the best overall interpretation and balancing of freestanding values. Set aside for now what it means to say one policy "best" satisfies those criteria; we turn to that topic in the next section. We have seen that many policies are reasonable or mutually acceptable. But Sara may only regard one policy as most reasonable. This section considers on what basis political liberals like Quong can say that Sara ought to favor the most reasonable policy, rather than selecting among reasonable policies using some other criteria. Call this the *problem of criterial indeterminacy*.

> *Criterial Indeterminacy*: Having narrowed down the space of feasible policy alternatives to a reasonable set, there remains the problem of selecting criteria for rank ordering reasonable policies to determine which to support.

Let me make two clarificatory points. First, the problem is one of rank ordering *feasible* policy, not of finding the best or most just policy, under idealized conditions. So, a solution to the problem need not involve offering a theory of ideal justice. Second, the problem concerns which policy citizens will *support*, either by voting for it or by defending it in conversation with their fellows.[16] Any reasonable policy enacted via democratic means is publicly justified, hence citizens have reason to endorse it *ex post*. As I use it here, endorsement is an attitude toward enacted policy. Support is an attitude toward policy that could be enacted in the future.

[15] Quong, *Liberalism without Perfection*, 206.

[16] Lister makes a similar distinction between *ex ante* and *ex post* reasons when he distinguishes public reason accounts that highlight reasons for decisions *ex ante* from those that justify coercion *ex post* (*Public Reason and Political Community,* 15–23).

A common conflation of two distinct civic duties obscures the problem of criterial indeterminacy for political liberals. Political liberals follow Rawls in saying that citizens have a "duty of civility." Only one of two common interpretations of that duty generates the problem of criterial indeterminacy. Some passages in Rawls emphasize that the duty of civility requires citizens to support the political positions they regard as "most reasonable."[17] Other passages say that the duty of civility only requires that citizens or their representatives support coercive policy that can be justified as reasonable.[18] There are two distinct duties at issue. To add clarity to the discussion, we will say that the *duty of civility* requires supporting the most reasonable policy, and the *duty of restraint* requires supporting some policy that passes through the filter of reasonableness. Political liberals follow Rawls in conflating these two duties, even though they are logically distinct.[19] Citizens can respect the duty of restraint while violating the duty of civility when they support a reasonable policy that they do not see as most reasonable. Civility requires criterial determinacy of citizens: they ought to always rank order policy according to what is most reasonable. Restraint does not require criterial determinacy: citizens can use any number of criteria to determine which reasonable policy to support.

It is easiest for political liberals to reject the duty of civility and embrace a lax interpretation of the duty of restraint. I call this position *agnosticism*, as it involves taking no stand on how citizens ought to rank order reasonable policy. The agnostic permits citizens to support any reasonable policy in the feasible set for any (morally permissible) reason, even when they regard some alternative as the most reasonable. Return to the case of Sara and Tony, and the two proposed policies that either require the church to employ women priests (L_1) or exempt the church from discriminatory hiring law (L_2). The duty of civility requires Sara to support L_1 because she regards it as most reasonable. In contrast, the lax interpretation of the duty of restraint permits Sara to endorse

[17] Rawls, *Political Liberalism*, 444.

[18] In one such passage, he says that the duty of civility requires that citizens show how the exercises of political power they "advocate and vote for can be supported by the political values of public reason" (Rawls, *Political Liberalism*, 217).

[19] Christopher Eberle's analysis of restraint is a clear exception to this generalization (*Religious Conviction in Liberal Politics*). Quong acknowledges the distinction between civility and restraint, without making much of it, in passages like the following: "the task of the political philosopher is to demonstrate that their theory is at least reasonable, or the most reasonable political conception possible" (*Liberalism without Perfection*, 226). Christie Hartley and Lori Watson are generally explicit that the duty of civility, on their view, requires appealing to the most reasonable political conception of justice (*Equal Citizenship and Public Reason*, 64, 82n49). Sometimes, though, they claim that civility only requires restraint (*Equal Citizenship and Public Reason*, 88, 136).

L_1 or to endorse L_2 because she recently converted to Catholicism and finds the tradition of a male-only clergy beautiful, even if not very liberal.[20] Additionally, the lax interpretation permits Sara to use a coin toss to decide between supporting L_1 and L_2, while entirely setting aside her considered judgment that L_1 is most reasonable. I take it that agnosticism is intuitively an odd view: surely, political liberals hold, Sara should support L_1 because she believes it is most reasonable. Yet political liberalism's methodological commitments seem to preclude saying that Sara ought to act on her judgment that L_1 is most reasonable.

Agnosticism is well motivated for political liberals who endorse what Quong calls the "buck-passing account" of justification. Theorists of political liberalism "pass the buck" on important justificatory questions onto citizens, who answer those questions for themselves by appeal to their full sectarian set of values and their controversial notions of truth.[21] It is much easier to pass the buck when justifying the lax duty of restraint, as opposed to justifying the duty of civility. For the former duty requires citizens find in their full set of values an answer to the question, "Why ought I to support some reasonable policy?" Comprehensive liberals may find their answer in an argument from the moral value of respect for persons.[22] Catholics may find their answer in the value of non-coercion promulgated in the Vatican II document, *Dignitatis Humanae*. Plausibly, very many diverse citizens can find unshared reasons to support some reasonable policy over unreasonable alternatives. It is less plausible that both comprehensive liberals and Catholics can find their own unshared reasons to support one specific reasonable policy over others in the reasonable set. So, "passing the buck" is much easier to do in a theory requiring that citizens embrace a lax duty of restraint, as opposed to the duty of civility.

Agnosticism is also well motivated by the view that shared values, and not specific conceptions of justice, are what justify public policy. Andrew Lister defends a "unanimous acceptability criterion" on which the only reasons that can justify public policy are those all reasonable citizens can accept.[23] Controversial interpretations, weightings, or applications of shared values fail the

[20] Quong is committed to the lax interpretation of the duty of restraint. He is explicit that reasonable citizens can be motivated by their comprehensive doctrine to support one reasonable policy over another (*Liberalism without Perfection*, 42). Because both policies fall within the reasonable set, citizens' justifying reasons are mutually acceptable, and that is all that reasonableness requires, in his view.

[21] For Quong, "the originality of Rawlsian political liberalism is that it delegates this task [of defending the truth of its theory] to reasonable citizens in a well-ordered society" (*Liberalism without Perfection*, 226).

[22] Larmore makes such arguments in "The Moral Basis of Political Liberalism."

[23] Lister, *Public Reason and Political Community*, 26.

test of unanimous acceptability. Consistent with unanimous acceptability, controversial interpretations and weightings can figure in a citizen's deliberations over which policy to support. But they would not be premises in the argument justifying that policy to other citizens. Such justificatory arguments would only include unanimously acceptable values and principles that admit of multiple interpretations and weightings.[24] This means many laws pass through the filter of unanimous acceptability, and they are all justified to reasonable citizens, despite those citizens substantively disagreeing over which laws that pass through that filter are best.

But these advantages all come with costs. Consistent with the lax interpretation of the duty of restraint, a citizen may find she has an all-things-considered reason not to support what she regards as the most reasonable policy. "Most reasonable" is a term of art. As I show in the next section, we can better understand the most reasonable policy as the one that, given a set of feasible alternatives, is required by what the citizen regards as the correct liberal conception of justice. Agnosticism bears the cost of permitting citizens to sometimes disregard their judgment of which policy, within a feasibly implementable space, is required by the correct liberal conception of justice. Not only will some citizens lack in their full sectarian set of values motivating reason to support what they take correct liberal justice to require, but it will also be the case that their judgments concerning correct liberal justice cannot appear as premises in the argument justifying a specific policy, as those judgments are not unanimously accepted among reasonable persons. So, agnosticism offers citizens nothing by way of shared reasons for determining what the most reasonable policy would be. Nor does it provide political liberals with much reason to carry on Rawls's project of theorizing about justice. Theories of justice, which offer correct interpretations and weightings of some of the shared basic values and principles about which reasonable citizens disagree, play no role in justifying public policy. Nor do they play any necessary role in guiding citizens' deliberations over how to answer the problem of criterial indeterminacy.

I do not consider here whether these costs are, all things considered, worth bearing. Instead, I describe what I take to be the main competitor to agnosticism: pluralism about justice. Pluralism holds that there are multiple conceptions of liberal justice that are correct, or "most reasonable." It is agnosticism's main competitor as a theory that assumes reasonable citizens can reasonably disagree even about very basic principles of justice. Pluralism's starting point is that some reasonable citizens can reject Rawls's claim, which we consider in detail below, that justice as fairness is the "most reasonable" conception.

[24] Lister, *Public Reason and Political Community*, 17.

Pluralism endorses criterial determinacy by holding that citizens should support the policy they see as justified by the correct conception of liberal justice. "Correctness"—which I use as a synonym for "most reasonable"—is always indexed to specific reasonable citizens, rather than to reasonable citizens as a whole. But "correctness" is not a subjective notion. Citizens determine which conception of justice is correct through intersubjective inquiry, including the method of reflective equilibrium. They intersubjectively determine which conception of freestanding liberal justice is correct by drawing on their full evaluative resources to order, interpret, and weigh materials in the shared evaluative framework. A liberal conception of justice, *a*, is correct as indexed to a specific constructivist device for comparing conceptions. And citizens determine which constructivist device is correct by drawing on their full evaluative resources. Political liberals should expect there to be multiple conceptions of justice that satisfy these correctness conditions, albeit for different citizens. Individual citizens see one unique conception of justice as correct. In contrast, theorists of political liberalism, with their commitment to epistemic abstinence, see there being a family of correct conceptions. This "family" is a proper subset of the broader family of reasonable liberal conceptions of justice, which Rawls discusses.[25] This proper subset is normatively distinctive in that it offers reasonable citizens *qua* citizens their answer to the problem of criterial indeterminacy.

3. Constructivist Correctness

Political liberals often attribute to citizens a duty that they will support the policy they regard as most reasonable. Yet no political liberal has offered an analysis of this concept. This omission makes sense if we assume all political liberals are agnostics who take no stand on which criteria citizens ought to use in rank ordering reasonable policy. Some of Rawls's remarks concerning the role of justice as fairness in political liberalism support reading him as an agnostic. In a letter to his editor, Rawls says justice as fairness has a "minor role" in political liberalism.[26] But he regards justice as fairness as the "most reasonable" member of the family of reasonable liberal conceptions of justice.[27] Rawls's arguments from *A Theory of Justice*, many of which he incorporates into *Political Liberalism*, justify him in assigning justice as fairness a privileged place among reasonable conceptions. Yet he says that his belief that justice as fairness has "a certain special place in the family of political conceptions" is "just an opinion of

[25] Rawls, *Political Liberalism*, xlvi–xlvii.

[26] Rawls, *Political Liberalism*, 439.

[27] Rawls, *Political Liberalism*, xlvi.

mine," which is "not basic to the ideas of political liberalism and public reason."[28] Presumably, what is basic is the notion of reasonableness. On the present analysis, one way of treating reasonableness as "basic" to political liberalism is to adopt agnosticism. Regardless of whether Rawls embraces agnosticism, there is a compelling way of reading him on which it is a mistake to say that his belief that justice as fairness is most reasonable is not basic to political liberalism. This is not because justice as fairness *per se* has a privileged place in the theory. Justice as fairness does not occupy a privileged place for all reasonable citizens, but only for a subset of citizens. The present analysis sees justice as fairness as privileged for citizens who view it as the correct conception of liberal justice, and therefore ought to use the criteria of justice as fairness for rank ordering reasonable policy. This section provides what Rawls did not: an analysis of the concept of "most reasonable." With this analysis, we can describe political liberalism's alternative to agnosticism.

Conceptions of justice, like justice as fairness, serve two functions: they specify how freestanding values are to be interpreted and ordered against each other in constructing arguments that defend policies as publicly justified. Those conceptions specify internal criteria for showing their interpretations and orderings of freestanding values to be better than alternatives. What Rawls, in "A Reply to Habermas," calls "pro tanto justification" achieves the first of these functions. There, political values are shown to be "suitably ordered, or balanced, so that those values alone give a reasonable answer by public reason to all or nearly all questions concerning constitutional essentials and matters of basic justice."[29] *Pro tanto* justification can only show that some conception of justice is reasonable, which, following Quong, I treat as passing through the filter of mutual acceptability. It cannot serve the function of rank ordering reasonable conceptions of justice to determine which is most reasonable. Elsewhere, though, Rawls offers remarks suggesting how to achieve this rank ordering. In "Lecture IV: The Idea of an Overlapping Consensus," Rawls, as I read him, refers to the family of reasonable liberal political conceptions as a "focal class" of liberal conceptions, and justice as fairness as "the center of the focal class" of liberal conceptions.[30] What makes justice as fairness "the center" for Rawls is that, among other conditions, "it is correctly based on more central fundamental ideas."[31] Justice as fairness is based on fundamental political ideas like society as

[28] Rawls, *Political Liberalism*, 451n27.

[29] Rawls, *Political Liberalism*, 386.

[30] Rawls, *Political Liberalism*, 167–68.

[31] Rawls, *Political Liberalism*, 168. Rawls adds a second condition: "it is stable in view of the interests that support it and are encouraged by it." Stability analyses require assuming a well-ordered society where everyone "accepts and knows that the others accept the same

a fair system of cooperation and the moral conception of persons. Its relation to these fundamental ideas is not deductive; rather, justice as fairness is connected to these fundamental ideas via the device of the original position.[32] In *Political Liberalism*, Rawls's original position argument is a device for "connecting" certain freestanding fundamental ideas in the public political culture with "definite principles of justice found in the tradition of moral philosophy."[33] In doing so, it serves the crucial function of rank ordering reasonable conceptions of justice, where the most preferred conception is most reasonable.

Rawls's original position models the choice situation of a hypothetical agent, where this choice situation "embodies" the various freestanding ideas found in the public political culture and requirements of practical reason.[34] These freestanding ideas include the "underlying conceptions of the person and of social cooperation" and "a particular understanding of freedom and equality."[35] This choice situation has four elements. An agent, who has (1) a well-defined utility function, confronts (2) a set of options from which she must chose. She is placed under certain (3) information constraints—she knows various things about herself, others, and the options she has to choose from—and her choice is guided by (4) principles of rational choice. Any problem of decision-making under uncertainty includes these four elements, so the original position models a problem of rational choice.[36] But this problem does not admit of a unique solution until we know more about the four elements. For Rawls, the content of elements 1 and 3 come from his interpretation of fundamental political values; the menu of options in 2 is given by the history of moral philosophy; and the arguments for 4 are primarily moral in nature, as I explain below. The original position argument, then, is thoroughly moralized, where the domain

principles of justice" (Rawls, *A Theory of Justice*, 4). It is not clear to me that political liberals should analyze the stability of a conception of justice in this sense. We are assuming that conceptions of justice guide an individual's *ex ante* deliberation, where this individual expects other citizens to be guided by rival conceptions. Surely, considerations of stability should figure in citizens' *ex ante* deliberation. But not all political liberals will agree that Rawls's idea of a well-ordered society should be central to stability analyses.

[32] Rawls claims that the original position "aims eventually to be strictly deductive," but this is not seen as an aim of the argument in his later work (*A Theory of Justice*, 104).

[33] Rawls, *Political Liberalism*, 339.

[34] Rawls, *Political Liberalism*, 90.

[35] Rawls, *Political Liberalism*, 339, 369.

[36] In *A Theory of Justice,* Rawls describes the original position argument as a part of rational choice. He later rejects this description, despite his original position having the formal features of decision-making under uncertainty. For the initial claim, see *A Theory of Justice*, 15. For Rawls's rejection, see *Political Liberalism*, 53n7.

of morality includes freestanding political values.[37] The argument purports to show that justice as fairness is most reasonable because a suitably constructed agent (with features 1 and 3), faced with a menu of principles of justice (2) to regulate her society, would select (informed by 4) justice as fairness over the alternatives after making pairwise comparisons among each.[38]

From the standpoint of political liberalism, Rawls's original argument is controversial. Not all reasonable citizens will accept it as the appropriate device for rank ordering reasonable conceptions of justice. This is because the argument, while presupposing only freestanding political values, nevertheless presumes various substantive interpretations of those values as correct. As an example, consider Rawls's claim that parties in the original position should employ maximin reasoning. Rawls famously gives a rational choice argument for this claim: according to his *argument from uncertainty*, maximin reasoning is uniquely rational in the original position's choice situation, given the informational constraints the chooser is under.[39] Some have objected that slightly different informational assumptions make it rational to employ different principles of choice.[40] The standard Rawlsian response to these objections is to turn to a new argument, the *argument from reciprocity*. According to this argument, cooperative schemes must be mutually beneficial and viewed by all participants—especially the worst off—as fair.[41] Maximin reasoning ensures that the principles selected guarantee everyone has a share of goods that they can live with, and that this share for the worst off is greater than what they would receive under the alternatives.[42] The argument from reciprocity is a moral argument,

[37] For Rawls's claim that freestanding political values are nevertheless moral, see *Political Liberalism*, 11n11.

[38] "We may suppose that this decision is arrived at by making a series of comparisons in pairs" (Rawls, *A Theory of Justice*, 106).

[39] Because the agent is behind the veil of ignorance, she is in a state of radical uncertainty: she does not know the probability of falling into the worst-off social class. She does know that the maximin strategy singles out an option she can live with if she finds herself, when the veil is removed, among the worst off. Additionally, the other options she could consider have worse outcomes that she could not accept. Interpreting the agent's utility function in this way, her choice of principles is a straightforward maximization problem.

[40] For an early statement of this objection, see Harsanyi, "Can the Maximin Principle Serve as a Basis of Morality?" For a more recent version of the critique, see Chung, "Rawls's Self-Defeat."

[41] Parties' utility function, Rawls says, "encodes certain basic features of our normative assumptions," including those about fairness (*Justice as Fairness*, 107). Although parties in the original position select a conception of fairness to regulate their society, the design of the original position also presupposes, and encodes, basic features of our notion of fairness.

[42] Also, reciprocity leads parties to derive minimal utility gains above a certain index out of concern that those gains may come at the expense of others who have less.

and it has justificatory priority over the argument from uncertainty in that it is what Rawls and his supporters appeal to when defending the rationality of maximin reasoning against utilitarian challenges.[43] That argument presumes a specific moralized notion of reciprocity, one that favors the difference principle over utilitarian alternatives.

What is Rawls's argument for accepting his specific moralized notion of reciprocity? Largely, his argument is the original position itself. This may make the original-position argument sound viciously circular. Given its function, it is not. Rawls acknowledges and intends that reasons for accepting his original-position argument eventually run out. The original position does not offer a deductive argument from truths known *a priori*. It takes certain things for granted. As Rawls puts it, "not everything, then, is constructed; we must have some material, as it were, from which to begin."[44] Starting materials include the interpretation of those fundamental ideas informing the construction of the original-position argument. Not all reasonable citizens will accept the same interpretation of these fundamental ideas. Rawls allows that "the public political culture is bound to contain different fundamental ideas that can be developed in different ways."[45] Certain ideas, when not presupposed in the process of construction, are clarified by that process. For example, speaking of the fundamental idea of respect for persons, Samuel Freeman notes that "so far as we aim to uncover the meaning of respect for persons for Rawls, it is explicated by justice as fairness."[46] For Rawls, the argument aims to proceed from "conditions . . . we do in fact accept," where "it helps us work out what we now think."[47]

As a political liberal, Rawls cannot help himself to the assumption that all reasonable citizens accept the conditions his original position argument presupposes. In this sense, he is correct to say that his belief that justice as fairness is most reasonable is not basic to political liberalism. But those citizens who justifiably reject some of the presuppositions of Rawls's original-position argument must accept something else in turn.[48] When they justifiably reject Rawls's notion of reciprocity, there must be some alternative notion of reciprocity they

[43] See, e.g., Freeman, *Rawls*, 194–97; c.f. Moehler, *Minimal Morality*, 82.

[44] Rawls, *Political Liberalism*, 104.

[45] Rawls, *Political Liberalism*, 227.

[46] Freeman, *Rawls*, 21.

[47] Rawls, *A Theory of Justice*, 514, and *Political Liberalism*, 26. See also Rawls, *Justice as Fairness*, 17.

[48] Neufeld and Watson make a similar point that anyone who rejects Rawls's original position argument must provide some compelling alternative showing why they regard their favored conception of justice as most reasonable ("The Tyranny—or the Democracy—of the Ideal?," 53).

accept. This alternative could even be that there is no uniquely correct notion of reciprocity in the public political culture. Whatever that alternative is, it provides materials from which a different device can be constructed to replace Rawls's original position. Like the original position, this new device would serve the function of rank ordering conceptions of justice. But the new device reaches its conclusions about how to rank order conceptions of justice on the basis of different presuppositions than Rawls's original position, raising the possibility that some alternative to justice as fairness would be the most preferred option. Justice as fairness is most preferred in Rawls's original position, making it most reasonable for Rawls, as the original position clarifies what he already thinks about justice. Another reasonable citizen, Alf, who justifiably accepts an alternative to the original position, may end up accepting some alternative conception of justice, *a*, as most reasonable. Of course, we cannot guarantee this would be the result: different devices for rank ordering conceptions of justice may yield the same verdict that justice as fairness (say) is most reasonable. Going forward, though, I will assume it is very likely that different devices would yield different conclusions about which conception of justice is most reasonable.[49]

So, Rawls and another reasonable citizen, Alf, regard different political conceptions of justice as most reasonable. I now argue they are correct in doing so. The ultimate standard of appeal for each citizen is reflective equilibrium. Reflective equilibrium is the standard by which Rawls assesses his interpretation of political values and the construction of his original position as a whole.[50] Rawls sometimes speaks of reflective equilibrium as achieved dialogically, from "the point of view of you and me."[51] This dialogue occurs between the theorist of political liberalism, who is constructing an original-position-style argument, and an individual citizen. Reflective equilibrium is achieved when the theorist's basic presuppositions—the things she takes for granted in constructing her choice model—match those of the citizen. It is not enough that the citizen recognizes the theorist's presuppositions as reasonable. The citizen must see those presuppositions as *correct* for the overall construction to be in reflective equilibrium for her. She does this by embedding those presuppositions in her

[49] This strikes me as the likely result because the conception of justice a device selects has a complex relationship to the device's starting assumptions. Those assumptions justify selecting that conception of justice. But the conception of justice also clarifies (or explicates) the starting assumptions. It seems to me unlikely that justice as fairness would do the best job clarifying a set of starting assumptions that includes the explicit rejection of Rawls's notion of reciprocity.

[50] Rawls, *Political Liberalism*, 70.

[51] Rawls, *Political Liberalism*, 28.

total set of values, showing, in what Rawls calls "full justification," that it is true she ought to endorse those presuppositions for the sake of determining what liberal justice requires.[52] Political liberals should expect that reasonable citizens will reach conflicting verdicts about which interpretations of shared political values are correct. So, political liberals should expect that there will be different devices, constructed out of freestanding political materials, for ranking conceptions of justice that are in reflective equilibrium for different reasonable citizens. If reflective equilibrium is the ultimate standard by which political liberals adjudicate the dispute between Rawls and Alf, they must accept the following conclusion: that both Rawls and Alf are correct in regarding distinct conceptions of justice as most reasonable.

Our rational reconstruction of why Rawls says justice as fairness is most reasonable has led us to acknowledging there can be a class of conceptions of justice that are most reasonable. Members of that class serve an important function for citizens by providing criteria for rank ordering reasonable policy. They do this by specifying an interpretation and weighing of freestanding political values—an *interpretation*, in short—that a specific citizen regards as most reasonable. Citizens appeal to this interpretation when determining which public policy proposal is most reasonable, out of a set of feasible, reasonable alternatives. Members of that class must also show why one specific interpretation is more reasonable than others. They do this via a ranking procedure, wherein a suitably constructed agent, who models relevant freestanding values, compares and rank orders different interpretations. The construction of this procedure presupposes as correct substantive interpretations of some (but not all) freestanding political values. It clarifies what a citizen now thinks on the presumption that she already accepts the device's presuppositions as correct.

A political conception of justice specifies an interpretation as most reasonable and includes some ranking procedure for comparing interpretations. A liberal conception of justice *a* is most reasonable (or correct) for a reasonable citizen Alf if and only if:

a. *a* specifies the correct interpretation, where this balancing is preferred to alternatives in a suitably constructed choice situation modeling ideas in the public political culture;
b. that choice situation correctly models the correct interpretation of ideas in the public political culture;
c. where both the model and the interpretation of ideas are in reflective equilibrium for Alf (or a suitable idealization thereof).

[52] Rawls, *Political Liberalism*, 386.

Although I have arrived at these conditions via a rational reconstruction of Rawls, the conditions should be acceptable to political liberals who reject many of the specific features of Rawls's original-position argument. Condition a leaves open the ultimate form the correct interpretation should take—whether this be as principles, à la justice as fairness, or as a series of trade-off functions, or as something else. Condition b does not require the choice situation take the same form as the original position, with a single agent, suitably constructed, representing all parties. Instead, condition b could be satisfied by a bargaining model with diverse agents representing parties.[53] Finally, condition c clarifies that the notion of correctness is indexed to a specific citizen. What is correct for Rawls need not be correct for Alf.

4. JUSTICE PLURALISM

Justice pluralism is the view that multiple conceptions of justice are most reasonable or correct. It gives the following answer to the problem of criterial indeterminacy. Citizens should support the policy they believe is most reasonable—that is, the policy justified by the interpretation of shared political values that, they justifiably believe, satisfies conditions a–c above. This is because, as citizens, they ought to support the policy they see as demanded by correct liberal justice. Determining which conception of liberal justice is correct is a complicated matter. Theorists of political liberalism must offer constructivist devices for rank ordering reasonable conceptions. Citizens must then determine which of these ranking devices is correct for them—using their full evaluative resources, including their controversial notions of truth—for ranking competing interpretations of some basic ideas in the public political culture that form the foundations out of which theorists build a device of construction. Even though the public political culture is shared, inquiry into that culture is marked by what Rawls calls the "burdens of judgment"—those same features of inquiry that, he believes, generate reasonable disagreement about metaphysical and religious matters under free institutions.[54] Citizens develop competing constructivist standards of correctness for interpreting and weighing the material of this common resource. Reasonableness can serve as a shared justificatory standard for evaluating enacted policy *ex post*.[55] But that standard

[53] For a recent bargaining model of the social contract, see Muldoon, *Social Contract Theory for a Diverse World*, 77–84.

[54] Cf. Rawls, *Political Liberalism*, 56–57.

[55] Quong distinguishes justificatory from foundational disagreement (*Liberalism without Perfection*, 193). On his analysis, disagreements about justice are *justificatory* because they presuppose shared standards of justification. Disagreements about religion and

fails to select determinate policies when citizens are deciding *ex ante* which to support. Instead, reasonable citizens deliberate in the framework of one family of most reasonable conceptions, where each member of this family represents a competing attempt to best make sense of the demands of liberal justice. This section considers a potential cost to pluralism: that it sees political conflict as an ineliminable feature in societies where all citizens honor the duty of civility.

Publicly justified policy can always be a source of conflict and opposition. For any given policy issue, there is a range of feasible, reasonable policies that could be implemented. Reasonable citizens have reason to endorse whichever member of that set is implemented. But publicly justified implemented law can fall far short of satisfying the standard of correct justice. Consistent with endorsing that law, citizens or their representatives can seek its repeal and replacement through legitimate means. These reasons for legitimate opposition do not necessarily go away when the law is most reasonable according to some conception. On the pluralist analysis, a law, *L*, is most reasonable if and only if:

1. *L* is selected as best out of a feasible set of alternatives according to the criteria of political conception of justice *α*, where
2. *α* is most reasonable for some reasonable citizen, Alf.

Alf's belief that *α* is most reasonable requires that *α* be in reflective equilibrium for Alf. But what is in reflective equilibrium for Alf need not be in reflective equilibrium for another reasonable citizen, Betty. Betty, who justifiably regards conception of justice *β* as most reasonable, can justifiably regard *L*'s implementation as a movement away from correct liberal justice. Moreover, Betty may justifiably harbor doubts that *L* is most reasonable for Alf. She cannot peer into his soul and perspicuously see *α* in reflective equilibrium. Reflective equilibrium is a function of a citizen's total belief set. Other things being equal, introspection grants a citizen better epistemic access to their own belief set than a peer could have. Yet that access is subject to distortions—say, from self-serving biases—making citizens blind to their own reflective disequilibria. Betty may rightly or wrongly be skeptical of Alf's claim that *α* is in reflective equilibrium for him. Either situation has a silver lining: her skepticism forces Alf to defend

comprehensive morality, in contrast, are *foundational* because participants do not share standards of justification for adjudicating their dispute. This distinction may make sense in the deliberative context of evaluating already enacted policy as publicly justified. But I doubt it always makes sense in the different context of determining which of a reasonable set of policies to support prior to enactment. For other criticisms of Quong's distinction, see Laborde, *Liberalism's Religion*, 99–110; and Vallier, "On Jonathan Quong's Sectarian Political Liberalism."

his beliefs about justice against old and new challenges, showing to himself and others that those beliefs are, indeed, in reflective equilibrium for him.

This generates what Rawls calls an "orderly contest" among rival conceptions of justice over time.[56] Presumably, Rawls calls these contests "orderly" because they proceed via democratically legitimate means. The "winners" are democratically enacted, but their victories may only be temporary, as any proposed or enacted law is subject to legitimate contestation. Citizens discuss with each other the merits or shortcomings of laws, voice opposition through protests to enacted law, and can seek out repeal through their choice of representatives. Contestatory politics can be heated and divisory, but there is an important sense in which the contest we are envisaging is different from sectarian disputes. We are imagining disputes among reasonable citizens who wish to honor the duty of civility. However heated these disputes may become, they are distinctively non-sectarian in that all participants share a commitment to correctly interpreting the public political culture. Their commitment has a practical foundation in the desire to treat one's fellow citizens as politically free and equal.[57] Regarding some specific policy issue, different citizens—Alf, Betty, and John Rawls—may all justifiably arrive at different conclusions regarding what such treatment entails.[58] But their shared desire to treat one another as politically free and equal leads them to seek out potential objects of overlapping consensus in the shared public political culture, elaborating that shared material (using unshared, controversial criteria) into concrete policy demands. Societies where all honor the duty of civility may be riven by conflict over which vision of justice to implement. Yet this is a conflict over how to best treat each other as politically free and equal.

Both agnosticism and pluralism see societies as divided by conflicts over which reasonable policy to implement. The key difference between the two is

[56] Rawls, *Political Liberalism*, 227. Rawls proceeds to say that this contest "is a reliable way to determine which one, if any, is most reasonable." In contrast, the present analysis sees this contest as a reliable way to determine which *are* most reasonable, and a legitimate way to determine which one is to be implemented.

[57] Weithman also argues that different conceptions of justice manifest concern with treating one's fellows as politically free and equal ("Autonomy and Disagreement about Justice").

[58] Neufeld and Watson offer a similar analysis, where reasonable citizens in a well-ordered society do not all endorse justice as fairness, but instead endorse a "reasonable" conception of justice ("The Tyranny—or the Democracy—of the Ideal?," 52–53). Yet Neufeld and Watson do not explicitly endorse pluralism, as they do not attribute to reasonable citizens a duty of civility to deliberate in the framework of what they regard as the *most* reasonable conception of justice. They deny that reasonable citizens would insist that society conform to their preferred conception of justice. On the pluralist analysis, reasonable citizens can nevertheless be deeply and justifiably dissatisfied with their regime when it fails to conform to what they regard as correct liberal justice.

that agnosticism permits citizens to pursue sectarian agendas within the space of reasonable policy. Pluralism, at least in principle, does not. Recall that agnostics endorse a lax interpretation of the duty of restraint, according to which a Catholic, say, can appeal to Catholic doctrine when rank ordering reasonable policy to determine which to support. Consistent with honoring the duty of restraint, a Catholic coalition can oppose subjecting the church to anti-discriminatory hiring law, all because there is a reasonable interpretation of religious freedom that permits them to do so. The agnostic holds that they can permissibly do this even when the members of that coalition justifiably regard an alternative interpretation of religious freedom, one that does not grant the church exemption from discriminatory hiring law, as most reasonable. This strikes me as a roundabout way of using politics to achieve sectarian aims. In the example, the Catholic coalition supports a reasonable policy for sectarian reasons. Of course, this is much better than supporting an openly sectarian policy for sectarian reasons. But it still poses a threat to the goods that public justification aims to achieve—specifically, that of civic friendship. Non-Catholics would have sufficient reason to endorse a law exempting the church from hiring laws. But they would, it seems to me rightfully, resent the members of the coalition who are motivated to support that law for sectarian reasons.

Unlike agnosticism, pluralism requires the Catholic coalition's motivating reasons for supporting a policy to be overdetermined. Consistent with honoring the duty of civility, members of this coalition may support a law both because they regard it as required by correct liberal justice and because they believe that it is, within the reasonable space, most consistent with Catholic dogma. Assuming it is common knowledge that they would not have supported the law had they not believed it required by correct liberal justice, the coalition's support manifests to others a concern for treating non-Catholics as politically free and equal. This common knowledge is difficult to achieve. Even when members of this coalition have the right motivating reasons, skeptical observers may see their support for the law as the Catholic tail wagging the politically liberal dog. So, the motivationally overdetermined Catholic may still threaten ties of civic friendship because non-Catholics cannot reliably distinguish her from the Catholic motivated by sectarian concerns. Nevertheless, there is an important conceptual distinction here that the pluralist can make and the agnostic cannot. The overdetermined Catholic manifests a virtue of civility that the agnostic cannot even acknowledge as a virtue. If you think this virtue is important, then you have reason to reject agnosticism in favor of pluralism.

Let me conclude by showing how pluralism is compatible with the method of epistemic abstinence. Pluralist political liberals ultimately pass the buck onto citizens to show some specific conception of justice is correct or most reasonable. Members of the resulting class of conceptions—the "family" of most reasonable conceptions of justice—are normatively distinctive in the following sense. Theorists hold that reasonable citizens who honor the duty of civility ought to deliberate in the framework of one of the members of this set. But a theorist cannot compare in the abstract, i.e., without referencing some specific citizen's full sectarian belief set, any two members of this family, α and β. The theorist cannot say of α and β that one is better or worse than the other.[59] In contrast, reasonable citizens can do this by referring to the criteria they determine to be correct in full justification. The epistemically abstinent theorist takes no stand on comparing the many correctness criteria that are in reflective equilibrium for different citizens. But it is important to note that only a proper subset of all reasonable conceptions of justice will be correct according to citizens. There will be some liberal conception of justice, γ, that no reasonable citizen regards as correct, leaving γ outside the family of most reasonable political conceptions. Consistent with practicing epistemic abstinence, pluralists can take a firm stand in saying that laws required by α or β are more just than laws required by γ. Laws required by γ might be publicly justified, and yet the theorist can consistently claim that society can do better, more closely approximating one of the visions of correct justice. In this sense, the pluralist theorist need not stand fully outside of what David Enoch calls the "political arena."[60] The pluralist theorist can occupy her impartial, epistemically abstinent high ground while condemning many reasonable laws as less than fully just. Only, this impartial high ground sees multiple conceptions of justice as correct. The theorist's impartial high ground is not the same one an individual citizen, Alf, occupies. Alf sees one specific conception of justice as uniquely correct given his endorsement of a specific controversial notion of correctness.

5. Conclusion

Whereas agnostics cannot avail themselves of any notion besides reasonableness in analyzing justified coercion, pluralists can appeal to reasonableness and the notion of a most reasonable conception of justice. A conception of justice narrows down the space of publicly justifiable policy to one unique option that is most reasonable, providing guidance for the citizen faced with the deliberative question of deciding which policy in that space to support. Pluralists see the family of conceptions of justice that are most reasonable, according to some

[59] It may also be that the theorist cannot claim that α and β are equally good. In that case, the theorist must treat members of the family as incommensurable with each other. For this definition of incommensurability, see Raz, *The Morality of Freedom*, 322.

[60] Enoch, "Against Public Reason," 134–36.

citizen, as isolating a special class of policy manifesting citizens' best attempts at treating each other as politically free and equal. Unlike agnosticism, pluralism offers an analysis of why reasonable citizens should honor the duty of civility. It also helps clarify what we can reasonably expect from peaceful political life in an ideal society where all honor the duty of civility. Political life should not be seen as a movement toward consensus on a single political conception of justice or as complacency with merely reasonable policy. Instead, it is a sphere of perpetual peaceful conflict among diverse visions of liberal justice, several of which the theorist of political liberalism can view as most reasonable.[61]

Center for Academic Pluralism, Heterodox Academy
gjesdal@heterodoxacademy.org

REFERENCES

Berlin, Isaiah. *The Proper Study of Mankind: An Anthology of Essays*. New York: Farrar, Straus, and Giroux, 2000.

Chung, Hun. "Rawls's Self-Defeat: A Formal Analysis." *Erkenntnis* 85, no. 5 (October 2020): 1169–97.

D'Agostino, Fred. *Free Public Reason: Making It up as We Go*. Oxford: Oxford: Oxford University Press, 1996.

Eberle, Christopher J. *Religious Conviction in Liberal Politics*. Cambridge: Cambridge University Press, 2002.

Enoch, David. "Against Public Reason." In *Oxford Studies in Political Philosophy*, vol. 1, edited by David Sobel, Peter Vallentyne, and Steven Wall, 112–42. Oxford: Oxford University Press, 2015.

Freeman, Samuel. *Rawls*. Abingdon, UK: Routledge, 2007.

Gaus, Gerald. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press, 2011.

Gaus, Gerald, and Chad Van Schoelandt. "Consensus on What? Convergence for What? Four Models of Political Liberalism." *Ethics* 128, no. 1 (October 2017): 145–72.

Harsanyi, John C. "Can the Maximin Principle Serve as a Basis for Morality? A

[61] I am immensely indebted to the late Jerry Gaus, who supervised my dissertation on which this paper builds. John Thrasher generously offered feedback on multiple drafts. For helpful discussion, I thank Jacob Barrett, Tom Christiano, Hun Chung, Keith Hankins, Brian Kogelmann, Ryan Muldoon, Greg Robson, Kirun Sankaran, Dave Schmidtz, Phil Smolenski, Kevin Vallier, Bas van der Vossen, Chad van Schoelandt, and Steve Wall. Finally, my gratitude to the anonymous reviewers whose criticisms improved the paper.

Critique of John Rawls's Theory." *American Political Science Review* 69, no. 2 (June 1975): 594–606.

Hartley, Christie, and Lori Watson. *Equal Citizenship and Public Reason: A Feminist Political Liberalism*. Oxford: Oxford University Press, 2018.

Laborde, Cécile. *Liberalism's Religion*. Cambridge, MA: Harvard University Press, 2017.

Larmore, Charles. "The Moral Basis of Political Liberalism." *Journal of Philosophy* 96, no. 12 (December 1999): 599–625.

Levy, Jacob T. *Rationalism, Pluralism, and Freedom*. Oxford: Oxford University Press, 2015.

Moehler, Michael. *Minimal Morality: A Multilevel Social Contract Theory*. Oxford: Oxford University Press, 2018.

Muldoon, Ryan. *Social Contract Theory for a Diverse World: Beyond Tolerance*. Abingdon, UK: Routledge, 2016.

Neufeld, Blain, and Lori Watson. "The Tyranny—or the Democracy—of the Ideal?" *Cosmos + Taxis* 5, no. 2 (2018): 47–61.

Quong, Jonathan. *Liberalism without Perfection*. Oxford: Oxford University Press, 2011.

Rawls, John. "The Idea of Overlapping Consensus." 1987. In *Collected Papers,* edited by Samuel Freeman, 421–48. Cambridge, MA: Harvard University Press, 1999.

———. *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press, 2001.

———. *Political Liberalism*. New York: Columbia University Press, 2005.

———. *A Theory of Justice*. Rev. ed. Cambridge, MA: Harvard University Press, 1999.

Raz, Joseph. *The Morality of Freedom*. Oxford: Oxford University Press, 1986.

Schwartzman, Micah. "The Sincerity of Public Reason." *Journal of Political Philosophy* 19, no 4 (December 2011): 375–98.

Sen, Amartya. *The Idea of Justice*. Cambridge, MA: Harvard University Press, 2009.

Vallier, Kevin. *Must Politics Be War? Restoring Our Trust in the Open Society*. Oxford: Oxford University Press, 2019.

———. "On Jonathan Quong's Sectarian Political Liberalism." *Criminal Law and Philosophy* 11, no. 1 (March 2017): 175–94.

Waldron, Jeremy. *Law and Disagreement*. Oxford: Oxford University Press, 1999.

Weithman, Paul. "Autonomy and Disagreement about Justice in Political Liberalism." *Ethics* 128, no. 1 (October 2017): 95–122.

Journal of Ethics and Social Philosophy
Vol. 26, No. 1 · November 2023

https://doi.org/10.26556/jesp.v26i1.2623

ARE SAVIOR SIBLINGS A SPECIAL CASE IN PROCREATIVE ETHICS?

Caleb Althorpe and Elizabeth Finneron-Burns

Hematopoietic stem cells are found in bone marrow and umbilical cord blood, and transplants offer sufferers of certain types of leukemia and anemia an excellent chance of surviving an otherwise terminal disease. However, stem-cell transplantation requires a donor who is a human leukocyte antigen (HLA) match to the recipient, and given the small size of modern bone marrow donor programs, the odds of a match are often minuscule. For instance, in the United States, there is a roughly 0.25 percent chance that an unrelated individual will be an acceptable match.[1] The odds of a match improve to 25 percent for siblings, since they both inherit the same HLA genes from their parents, but due to the average size of the modern family in the West, sufferers of leukemia and anemia will usually lack an existing sibling match. As a result, parents of children suffering from these diseases may wish to conceive a child to provide the necessary stem cells from the newborn's umbilical cord blood to save the life of their existing child. Modern technology means that parents do not need to just conceive and hope to hit the HLA jackpot. Rather, they can use in vitro fertilization (IVF) to produce multiple embryos and follow up with preimplantation genetic diagnosis (PGD) to select one or more that are an HLA match for the sick child.[2] At birth, the cord blood is collected from the umbilical cord and transplanted to the sick child. This is curative in up to 90 percent of noncancerous patients and has a five-year survival rate (the benchmark for cancer remission) of at least 68 percent of patients with leukemia.[3] In both cases, the chances of survival are more than doubled by using a related rather than unrelated donor.

Children conceived in order to donate the stem cells in their cord blood are examples of "savior siblings," a term referring to children intentionally created

[1] Robertson et al., "Conception to Obtain Hematopoietic Stem Cells," 35.

[2] PGD is the genetic profiling of fertilized embryos for certain characteristics (such as HLA type or inherited conditions such as Huntington's disease) before they are implanted.

[3] Leung et al., "High Success Rate of Hematopoietic Cell Transplantation Regardless of Donor Source in Children with Very High-Risk Leukemia."

to donate biological material, most commonly cord blood, but theoretically also bone marrow or solid organs (liver and kidney), to save the life of an already-existing child. Some writers in medical ethics have argued that there are features inherent in the creation of savior siblings that make the practice impermissible or should at least make us skeptical about the arguments offered in its favor.[4] The primary reasons that have been offered against the practice are: (1) creating a savior sibling has negative impacts on the created child, and (2) creating a savior child represents a wrongful procreative motivation of the parents. In this paper, we examine the extent to which the creation of savior siblings actually presents a special case in procreative ethics. We do not deny that there is a unique feature present in the savior sibling case—namely, that the child was created to save their sibling's life. We also do not claim that this unique feature raises no novel normative questions for procreative ethics (e.g., whether there are any conditions under which the creation of savior siblings might be morally obligatory). But what we *do* deny is that the distinctive feature of *being a savior sibling* is what makes the procreative act wrong. Our conclusion is that what would make the creation of a particular savior sibling permissible or impermissible are the same things that would make the creation of any child permissible or impermissible. Our conclusion is that savior siblings—in relation to the reasons for the permissibility or impermissibility of their creation—are not a special case in procreative ethics.

There are two clarificatory points to make at the outset. First, our discussion relates to savior siblings created to donate umbilical cord blood, bone marrow, and/or solid organs. However, due both to continual improvements in the efficacy of cord blood transplants and the availability of a cord blood donation at the time of birth, we take cord blood donation to be the prototypical savior sibling case.[5]

[4] Wolf et al., "Using Preimplantation Genetic Diagnosis to Create a Stem Cell Donor"; Chan and Tipoe, "The Policy Statement of the American Academy of Pediatrics."

[5] To elaborate, multiple studies have found that sibling-matched cord blood transplantation can be just as effective in treating blood diseases/cancers as bone marrow transplants, with possibly fewer complications. See Rocha et al., "Graft-Versus-Host Disease in Children Who Have Received a Cord-Blood or Bone Marrow Transplant from an HLA-Identical Sibling"; Bizzetto et al., "Outcomes after Related and Unrelated Umbilical Cord Blood Transplantation for Hereditary Bone Marrow Failure Syndromes Other Than Fanconi Anemia"; Locatelli, "Outcome of Patients with Hemoglobinopathies Given Either Cord Blood or Bone Marrow Transplantation from an HLA-Identical Sibling." And so, given that the median time between beginning the first cycle of IVF-PGD and the birth of a savior sibling is 3.7 years, and that young babies cannot donate bone marrow, it is expected that a cord blood donation would be performed in the first instance, with the possibility of needing a bone marrow donation later if the cord blood transplant is unsuccessful. See Kakourou et al., "Pre-Implantation HLA Matching," 80–81. Thus, savior siblings would not

Second, in order to determine whether or not savior siblings are "special," we need to know what the relevant comparand is—special compared to what? Since we are interested in whether or not savior siblings present a special case in *procreative* ethics, the relevant comparand is what we will call the *Standard Child*. This is a child created (in part) for any number of nonsavior reasons, such as the parents' desire to have additional children, to provide companionship to their existing children, to please grandparents, and so on (more on these reasons in section 2). It is our contention that being a savior sibling does not raise special normative concerns relative to the Standard Child. Finally, to ensure we do not stack the case in our favor, in the savior sibling case we will have in mind parents who would not otherwise have chosen to have an additional child.

1. NEGATIVE IMPACTS ON THE SAVIOR SIBLING

1.1. Physical Harm

The first and most obvious reason to consider savior siblings a special case in procreative ethics is the notion that their creation harms the created child and that such harm is not present in the creation of nonsavior siblings. Such harm, the argument might go, could be sufficiently serious that it makes the creation of a savior sibling unjustified regardless of any benefits it might afford the sick child.

The first point to make is that the special harm cannot be from the use of IVF and PGD in the selection process itself, as these treatments are not unique to the savior sibling case. IVF is used around the world by infertile and LGBT couples, and PGD is available to families with a history of inherited diseases such as cystic fibrosis and Huntington's disease to select for children who will not suffer from these serious and often fatal conditions. If there is any harm associated with IVF and/or PGD *per se*, there is a harm to *all* embryos created or selected in this way.

When it comes to physical harm *after* birth, we will first consider cord blood donation.[6] The case of cord blood donation is simple, because this procedure

be created to *be bone marrow donors* but to be cord blood donors, with the knowledge that there may be bone marrow donation at a later date. And the same point applies to organ donation, since given that living solid organ donation is only ever ethically (and legally) permissible with the patient's informed consent, any permissible organ donation by a savior sibling could only ever occur many years after their creation. This means that in the savior sibling case, bone marrow and solid organ donations, unlike cord blood donation, are only possible, not inevitable scenarios.

[6] Of course, there are those who believe that *any* procreation harms the resulting child since every life will inevitably include some elements of pain or suffering, such as David Benatar (*Better Never to Have Been*). But this is also not special to the savior sibling case.

results in no physical harm at all. The collection of cord blood is noninvasive and painless (it is taken from the placenta after it is delivered), and studies have shown that collecting this blood poses no risk to the newborn.[7] And so, while this brings in a difference (not all children have their cord blood collected, although the vast majority could), it is not a difference that is relevant to a claim that the creation of savior siblings is a special case in procreative ethics.[8] Indeed, thousands of mothers, including one of the authors, voluntarily donate their newborns' cord blood to strangers via public blood banks every year, and others collect and store it in case it is needed by their own child in the future.

However, if the savior sibling case is one where bone marrow or solid organs end up being donated, then physical harm will occur, as these donations are more physically invasive and, in the case of solid organs, can require significant recovery time. While this might make the savior sibling case initially appear very different from the Standard Child procreation case, that would be too quick. This is because any tissue donation is only ethically permissible under certain conditions, and these conditions apply just as much to a savior child as to the Standard Child.

For example, the American Association of Pediatrics (AAP) has laid out criteria under which it is ethical for precompetent minors to donate bone marrow, and these criteria would need to be met regardless of whether the child was created as a savior sibling or not.[9] Similarly, any organ donation (and its affiliated harms) is only ever ethically permissible when the patient has given informed consent.[10] These requirements will apply just as much to the savior sibling case as they will to the Standard Child—they are not trumped by a savior sibling's reason for genesis.

[7] Rubeis and Steger, "Saving Whom?," 480–81.

[8] Mother-baby dyads who cannot give cord blood include those with inherited medical conditions or infectious diseases, babies conceived by donor eggs/sperm, twins/triplets, and babies born more than six weeks prematurely.

[9] The criteria are: (1) no adult matches are available; (2) there is a strong, positive relationship between the donor and recipient; (3) there is some likelihood that the recipient will benefit from the transplant; (4) the risks to the donor are minimized and reasonable in relation to the benefits accrued to the donor and recipient; (5) parental (and sometimes donor) consent is obtained. See Committee on Bioethics, "Children as Hematopoietic Stem Cell Donors."

[10] Richards, "A World of Transferable Parts," 381; Saunders, "Consent and Organ Donation," 312–13. In addition, almost all legal jurisdictions have lower age limits on living organ donors, usually sixteen years. Where there are no limits (e.g., England), there is still a requirement for informed consent. This suggests that what would be concerning in the organ-donation case is whether being created as a savior sibling influenced informed consent. We consider this below (section 1.3).

One might think, however, that there is still a difference in the sense that the savior child is being created *in order* to be a bone marrow or organ donor and therefore experiences pain. To this we have two replies. The first points out that it is not actually the case that the child *is* being created in order to donate and experience pain. For one thing, they are being created in order to save their sibling. But given cord blood is the only *inevitable* form of donation (see note 5 above), a better formulation of the objection is that they are being created *with the knowledge* that they may later donate and therefore experience pain or harm.

Yet this possibility that the child will experience pain/harm is not enough to establish that savior siblings are special. Any time you choose to create a child, you do so in the knowledge that they may be harmed and/or experience pain in the future. The question is whether there is anything special about creating a child you know may experience pain *in this particular way*—namely, as a future bone marrow or organ donor—as opposed to a multitude of other ways (car crashes, sports injuries, broken hearts). One difference might be that in the tissue donation case, the pain or harm will be experienced for the benefit of another. But this can be true for the Standard Child too. He could be injured in a car accident on the way to take his sibling to a sports practice or dentist appointment. Another difference might be that the pain or harm is not totally random but perhaps reasonably foreseeable. However, this could also be true for a Standard Child who is created and strongly directed by her parents to play sports such as ice hockey, rugby, or horseback riding that have a high probability of sometimes serious injury.

1.2. Psychological Harm

Perhaps what makes the savior sibling case special is that it leads to psychological harm for the created child that would not be present in the Standard Child case. Of course, it is impossible to predict the psychological well-being of *any* child before conception, and this is equally true of savior siblings. But in both the savior sibling and Standard Child cases, two features seem relevant to any prediction about psychological harm: the child's being told why they were conceived and the nature of the parents' attitudes to or treatment of the child. We consider each in turn.

Let us assume that a savior child is informed of their reason for genesis. They will grow up knowing they were created in order to save their sibling. What might the psychological impact of that be? Critics might think it could lead the child to fail to see herself as a person with dignity who is worthy of respect. The child may feel as though they were not *really* wanted by their parents or that their parents took on an unwanted burden by having them. But similar sentiments could also occur in many forms of the Standard Child case. Take, for example, children who result from a contraceptive failure or sexual assault. Upon learning the cause of their creation, these children will surely be just as (if not more) likely as savior siblings to feel they were not *really* wanted by their parents. But the point here extends beyond only "unplanned" cases of procreation. This is because, as we will outline in more detail in section 2, it is not only conceptually impossible to create a child purely for their own sake (because no specific child exists at the point of conception) but also morally undesirable to do so (because some instrumental value is key to positive personal relationships). This means, in principle, that there is always the potential for a child, upon coming to learn the reason for their existence, to feel as if *they* were not *really* wanted by their parents (say, because they were created to give an older child a sibling, or to pass on genes, or due to personal fulfillment from parenting).

However, a different worry about psychological harm could be that the savior child might feel like a failure and have low self-esteem if their donation does not save their sibling, who dies despite their donation (even though this is unlikely due to the very high success rates), and that no such potential sense of failure is possible in the Standard Child case. But an individual's self-esteem or any lack thereof cannot be read directly off how they fare against some standard taken as important by third parties. This is because self-esteem, as a self-regarding attitude, depends on a person's *own* beliefs about what standards are important and how they fare against them.[11] As such, insofar as it is possible for children in the Standard Child case to believe that they failed to live up to their parents' expectations or act in a way consistent with the standards related to the reason for their creation (which, to them, the meeting of which would likely be very important), a similar concern with a sense of failure and that they disappointed their parents applies just as much in their case too. This is so even though, to others, the standard might seem less important than the one in the savior sibling case. Taking one of the common reasons for procreation just mentioned, if a child knew that the reason for their existence was to give their older sibling a friend, but ended up being disliked by their older sibling, then possible psychological harm resulting from hits to their self-esteem and letting down their parents seems just as likely.

Of course, it is undeniable that a savior sibling might experience psychological harm not only from being informed of the reason for their existence but also due to the way they are treated by their parents. Examples might include parents treating them as an unwanted burden, regularly reminding them that they did not really want them, or visibly favoring the older child. But these poor

[11] Sachs, "How to Distinguish Self-Respect from Self-Esteem"; Dillon, "Self-Respect and Self-Esteem."

parental behaviors could sadly happen to *any* child. Children in the Standard Child case can be subjected to similar treatment, such as being told they were an "accident" or that they are a burden and make their parents' lives so much harder, and so on. It seems to us that in both of these cases, the wrong lies in the parents having related to their children (savior or standard) in negative ways or taken certain negative attitudes toward them. Any cause of psychological harm to the savior child is not that they were created to be a savior sibling as such but instead the negative way or ways in which the parents relate to the child as they grow up. Whether *any* child feels loved and grows to see themselves as a valuable end in themselves depends substantially on how they are treated by their parents; savior siblings are not unique in this respect. The same point applies to other kinds of nonphysical harm that might potentially be experienced by the savior sibling, such as receiving less material benefit or fewer opportunities relative to the older child. Such harms are contingent on the actions of the parents; they are not inherent in being a savior sibling.

However, one might object that although these potential psychological harms can happen to any child, they are more likely to happen to savior siblings than other children. We disagree. Although there is always the possibility that parents will treat their savior child badly (which, as we have said, is unfortunately a possibility for any child), we think such treatment is at worst equally likely and at best much *less* likely to be experienced by a savior child. It would be a strange person indeed who cared so deeply about their first child that they were willing to conceive, bear, and raise another child to save them, yet also be so callous and unloving toward the second, savior child to whom they stand in exactly the same biological and parental relationship they do to the first. Indeed, in studies investigating the attitudes of parents who decided to create a savior sibling, parents flatly rejected the idea that any person who was willing to go through IVF and PGD could then mistreat the resulting child or treat them differently from the older child.[12] If anything, then, it seems more likely that parents of savior siblings will treat them in a way that makes them feel like a *hero* because they did something no one else could do—save their brother or sister. These children may well embrace their identity as a savior sibling as a badge of honor.

1.3. Violations of Autonomy

The next potential reason savior siblings might present a special case in procreative ethics is that their creation violates the savior child's autonomy. If violation of autonomy is understood to occur when things are done to a person

[12] Strong et al., "It's Time to Reframe the Savior Sibling Debate," 19–20; Haude et al., "Factors Influencing the Decision-Making Process and Long-Term Interpersonal Outcomes for Parents Who Undergo Preimplantation Genetic Diagnosis for Fanconi Anemia," 651.

without their consent, then perhaps creating a savior sibling to donate stem cells violates the savior sibling's autonomy.[13] In these cases, the donor child is far too young to grant or withhold their consent to the donation and the procedure relies on the consent of the parent. In medical ethics this is known as the stage of precompetence. Some critics move directly from the inability of a savior sibling to give consent to the procedure to a claim that, as a result, the procedure constitutes a direct violation of the child's autonomy in a way that counts against the permissibility of the practice.[14]

In relation to cord blood, the first point to make is that the mere fact a savior sibling does not consent to the donation of cord blood cannot serve as an argument that savior siblings are special, because it mirrors the uncontroversial and not uncommon practice in the Standard Child case of parents choosing to donate their newborns' cord blood to public cord blood banks.[15] This aside, it is doubtful that on its own, the lack of consent to an action that causes no physical harm and no increased likelihood of psychological harm (such as cord blood donation) is sufficient to constitute a violation of the autonomy of a precompetent child. Seeing simple lack of consent as a violation of autonomy would commit one to regarding virtually all actions toward precompetent children, including completely innocuous ones, as violations of their autonomy. But this cannot be right. You do not violate your precompetent child's autonomy when you change their nappy or take them with you to the supermarket without their consent.[16] With regard to bone marrow donation, there is no autonomy-related difference between savior children donating and the Standard Child donating. In both cases, certain criteria need to be met.[17] If donating bone marrow

[13] This point does not apply to organ-donor savior siblings because organ donations ethically and legally require a person's informed consent.

[14] Rubeis and Steger, "Saving Whom?," 480; Chan and Tipoe, "The Policy Statement of the American Academy of Pediatrics," 3.

[15] One objection here might be that a relevant difference is that parents of savior sibling are likely to be biased when it comes to a decision about donation, given that is the very reason why they decided to have the child in the first place. But if parental bias were a relevant difference (which we are not sure is likely—see above), it is more of a concern with using parental consent as a proxy for a child's best interests (which also occurs in the Standard Child case), not a concern that the creation of savior siblings and subsequent donation cannot be in the interests of the child.

[16] This does not deny arguments claiming children (even young children) possess the capacities relevant to autonomy. See Mullin, "Children, Autonomy, and Care"; Hannan, "Childhood and Autonomy," 115–18. The savior sibling case relevant here (cord blood donation) concerns actions toward children that are precompetent infants. We are not aware of any argument that claims these children possess autonomy.

[17] At the moment, the accepted criteria are those of the AAP; see note 9 above.

wrongfully violates a child's autonomy, then that is wrong whether that child was a savior sibling or not.

An alternative objection might be that savior siblings do not consent to be created for the purpose of their donation. Might this be a relevant difference from the Standard Child case? Not if consent is understood as express consent, due to the simple fact that nobody consents to the reasons for their own creation. The argument might, however, be put in terms of hypothetical consent and go something like the following: while I might hypothetically consent to be created for the array of purposes that make up the Standard Child case, I would not hypothetically consent to be created for the purpose of donating my stem cells to save my sick older sibling. This is plausible, but it is not clear that the concern here is still with autonomy. This is because to make a statement about hypothetical consent, we need to talk of the *reasons* why such consent would or would not hypothetically be given. But once we are talking of reasons, it is unclear what work hypothetical consent is actually doing in the argument. If it is some reason *x* that makes us say that a person would not hypothetically consent to be created as a savior sibling, then it is *that* reason that provides the argument against the practice: a person's hypothetical consent provides no independent argument.[18] To claim savior siblings are a special case due to a lack of hypothetical consent, then, is not to claim that creating savior siblings is special due to its effects on the created child's autonomy but that it is special for some other reason.

Perhaps creating savior siblings is normatively different from the Standard Child case because it affects the autonomy of the child in the future. For instance, Matthew Clayton argues that actions toward a minor violate their autonomy if, once they have reached a stage of competence, they would denounce the treatment. That is, what counts is *retrospective* consent. The cases that Clayton thinks are problematic are those that can be seen as deciding for the child the goals they will pursue later in life.[19] Similar concerns underlie the claim that children have a right to an "open future," which requires key options to remain open until a child is a self-determining adult who can choose among them. An example Joel Feinberg gives is Amish parents refusing to send their children to public schools,

[18] As Ronald Dworkin puts it: "Hypothetical contracts do not supply an independent argument for the fairness of enforcing their terms. A hypothetical contract is not simply a pale form of an actual contract; it is no contract at all" ("The Original Position," 17–18). David Enoch has recently argued that hypothetical consent might be normatively significant and be connected to the value of autonomy in situations where it better respects a person's deep and central commitments (Enoch, "Hypothetical Consent and the Value(s) of Autonomy"). As we are about to argue, we do not think the savior sibling case connects to autonomy so understood.

[19] Clayton, "The Case against the Comprehensive Enrolment of Children."

given this drastically limits the occupational choices that will be available to them.[20] This focus on a child's future autonomy is more appealing than making a wholesale claim about all actions in the absence of informed consent being violations of autonomy, as it explains why certain innocuous actions are not violations of autonomy (taking your precompetent child to the supermarket) while others plausibly could be (taking your precompetent child to be baptized).

Does the creation of a savior sibling to donate cord blood and possibly bone marrow fail to respect a child's future autonomy in a way the creation of the Standard Child does not? We think not. The medical procedure does not lock a child into a particular way of life before they get the chance to choose for themselves, nor does it close off a set of key options. From the perspective of future autonomy, being created for the purpose of stem-cell donation is less like the actions Clayton and Feinberg find problematic and more akin to the multitude of actions that are permissible to do to a precompetent child unable to give their consent, such as choosing their hairstyle, giving them a well-tested vaccine, or taking them with you on a car ride. Just like haircuts, vaccines, and car trips, donating stem cells does not fix the options or limit the horizons available to children once they become competent to choose for themselves.

However, an argument might be made that creating a savior sibling and donating their cord blood *is* changing the likelihood of a child's future choices in one important respect—future donations (including more invasive or permanent donations to their sick sibling such as solid organs such as the kidney or liver). The worry is that a savior sibling may be pressured later in life to donate again if their sibling relapses or develops new illnesses and that this pressure would constitute a violation of autonomy (a similar concern will apply to any child who is the candidate to save a loved one's life through donation). This concern, then, overlaps with the worry briefly signaled earlier (in note 10) regarding the organ donation savior sibling case and the thought that being created for this purpose might affect informed consent by putting undue pressure on the child (who has reached a stage of competence) to donate.[21]

A common view in bioethics is that for a patient to give their informed consent to a medical procedure, they need both to have an adequate understanding of the risks involved and, perhaps more relevant to the discussion here, to be

[20] Feinberg, "The Child's Right to an Open Future," 77, 81–82.

[21] Once a child has reached a stage of competence, it would be wrongfully arbitrary to treat their voluntary decisions regarding donations any differently from voluntary decisions of competent adults. Of course, the safeguards needed to ensure the decision to donate is in fact voluntary might still differ between competent children and adults. See Wilkinson, *Ethics and the Acquisition of Organs*, 138–44; Brierley and Larcher, "Organ Donation from Children," 1178.

free from any kind of "controlling interference," where controlling interference is understood as the active intervention by other agents.[22] Similar accounts put normative weight on informed consent because it ensures patients are able to make self-authored decisions that are the result of their own judgment and reflection, making them responsible for the shape of their own life.[23]

Three barriers to this aspect of informed consent are commonly identified: coercion, undue inducement (i.e., positive offers clouding rational judgment), and "no-choice" situations.[24] Is a savior sibling's choice to donate an organ particularly vulnerable to any such barriers? If these barriers occur due to problematic behavior and actions of parents, then it will, of course, fail to be a case of informed consent. But this just becomes another instantiation of the nonphysical harm case. It is, of course, possible for a savior sibling to feel pressured to donate because of threats from their parents, but coercion and threats (even implicit ones) to undertake serious life decisions (such as donating a solid organ) are serious parental wrongs that violate a child's autonomy regardless of whether a child is a savior sibling. So are actions aiming to "nudge" a child into making one decision over another (say, by intentionally influencing a child's decision about donation by beginning and ending all conversations with how fantastic it would be if they donated) or framing the issue as one that makes the child feel as if there is really only one choice to be made ("Once you have undergone the donation …").

What we take to be the more serious charge is the concern that even if a savior sibling is raised in a loving environment that inculcates in them a strong confidence in their individual worth, the mere fact that they are a savior sibling might put undue pressure on them that influences their decision to donate an organ. Perhaps this fact is enough to make the decision to donate an organ to save a loved one appear to be a "no-choice" situation. The thought is that because the choice not to donate is such a horrible alternative (as it results in one's sibling dying), the voluntariness of the choice has been undermined. But as Nir Eyal outlines, we need to distinguish between cases where the curtailment of options results from the offer itself and those where the curtailment is merely the result of (often nonideal) circumstances, as it is often only the former that appears to undermine the voluntariness of a decision.[25] Applied to the case of concern here, the relevant distinction is the difference between giving the child the option

[22] Beauchamp and Childress, *Principles of Biomedical Ethics*, 100–101.

[23] Dworkin, *The Theory and Practice of Autonomy*, 108–20.

[24] Eyal, "Informed Consent"; Campbell et al., "How Young Is Too Young to Be a Living Donor?"

[25] Eyal, "Informed Consent."

of donating an organ or having their sibling murdered and giving the child the option to donate or not, where the latter will result in their sibling dying from disease. The choice of a savior sibling not to donate does, of course, have such a serious downside that it would make such a decision unlikely, but this is no different from other scenarios involving medical procedures. The voluntariness of a crash victim's decision to consent to a lifesaving leg amputation need not be undermined just because the alternative (they die) is horrific.

Our argument, then, does not deny that features affiliated with being a savior sibling might influence a child's likelihood of consenting to a donation. That seems undeniably plausible. Our argument denies that changing the likelihood of future decisions makes the savior sibling case special relative to the Standard Child case, because features affiliated with the Standard Child case will influence important future decisions too—and so this fact alone is not enough to support the position that the savior sibling case is special. We think it is the possible influence of two kinds of features of procreative cases that are relevant here: the influence on future choices exerted by a child's knowledge about the reason or reasons for their existence and the influence on future choices exerted by environmental conditions. In the savior sibling case, the former feature would be the child's knowledge that the reason for their existence is in part to provide biological material to save the life of their older sibling, while the latter would be the fact that they have the right genetic profile that makes a donation to their sick sibling possible.

The way such features might influence a savior sibling's choice to donate is clear enough. But take, for instance, the following uncontroversial example of the Standard Child case: two persons whose reason for having a child is in part to have an extra person around to share their love of music. This child will, first, grow up knowing that the reason for their existence is in part their parents' wish to share their love of music; second, the child will be raised in a "music-dense" environment (their parents are always playing music, discussing it, and putting up band posters around the house). Both these features will surely strongly influence the likelihood of the child's choices regarding nontrivial life options (such as what career they choose to pursue, their choice of a partner, and so on), all in a way that we think is analogous to the influences on the decision of a savior sibling to donate an organ.

An objector might reply here that two relevant differences remain between the savior sibling case and Standard Child cases that make any influence over a decision to donate particularly concerning. First, the choice to donate an organ involves physical harm and risks, and second, the stakes of the decision are very high. But some examples show that harm and high stakes are often also present in the decisions that features of Standard Child cases influence.

Regarding harm and risk, a person's choice to earn their living as a musician instead of a more secure career is a choice to undergo psychological stress and anxiety; a person's choice to play football as a hobby over chess is a choice to expose themselves to higher chances of concussion. And as we have argued, parents' motivation for procreation and the environment in which the child is raised often influence the likelihoods of these kinds of choices.

Now, the stakes involved in a decision to donate are obviously extremely high (do it or my sibling dies). But we cannot underplay here the high stakes in other decisions that often present themselves to children in Standard Child cases. We think the best example here is decisions regarding the endorsement of religious and other comprehensive beliefs. When parents who share a religious belief decide to have a child, they are in effect (knowingly or not) putting the child in a position where in the future they will face the following high-stakes decision: endorse or remain a follower of a particular religious belief, or no longer share their family members' conception of the good. The stakes involved in this decision need not result from any malice or pressure from parents (there is no threat of disownment) but simply the fact that a certain distance is unavoidably introduced between persons who do not share the same comprehensive conception. To not be able to fully understand family members' outlook on the world, appreciate their moral compass, or have deep conversations about what they hold most dear are all great losses.

Consequently, the decision of parents in the savior sibling case to put their child in a situation where they will (potentially—see note 5 above) have to make a decision that involves both harm and extremely high stakes does not make the parents of savior siblings unique. In commonplace Standard Child cases, both the reason for the child's existence and the environment in which they are raised can also influence the likelihood of a child's choice in decisions with the same features. Such influences are inevitable given the social contexts in which persons make decisions, and so long as parents in these cases do not explicitly pressure their child and ensure a range of different alternative options is available and known to them, there is no violation of their autonomy.

2. Wrongful Parental Motivations, or Instrumentalizing the Savior Sibling

The creation of savior siblings might also be thought to be a special case for nonconsequentialist reasons. The claim that seems most relevant here is the allegation that the creation of a savior sibling wrongfully treats the child instrumentally. If this were the case, then it would certainly be a reason to object to creating savior siblings.

Worries of this kind are common in the medical and bioethics literature. For instance, Lord Robert Winston—a pioneer in fertility technology—argues that creating a savior sibling "would be using an unborn child as a commodity."[26] We think there is an intuitive force to this objection and that it would apply to all forms (cord blood, bone marrow, and organ donation) of the savior sibling case. As such, we disagree with the way Sally Sheldon and Stephen Wilkinson portray the nature of the burden of proof as one where an objector to the practice of savior siblings "must demonstrate that these [sick] children's deaths are less terrible than the consequences of allowing this particular use of [PGD]."[27] This agent-neutral outlook misses the deontological concern with instrumentalizing others. The reason we have assumed that the parents would not have otherwise chosen to have another child is to put the concern with instrumentalization in its strongest terms. However, it is necessary to look more closely at this claim.

A very common reason parents with more than one child give for having had their second child is to ensure that their first child has a sibling—to play with, to have as support when older, to have help taking care of aging parents down the line, and so on. We will call this the *Companion* case. If this is correct, then parents in these situations are creating the second child at least partly for the benefit of the first. Rarely if ever do we encounter moral criticism of such parents. Rather, they are often lauded for taking on some costs (those of raising another child) for the benefit of their other child. Having a second child to give a first child a sibling is at worst considered morally neutral. The mere fact, then, that in the case of savior siblings the second child is created to benefit the first child will not be enough to sustain the claim that savior siblings are special in this regard. If it is permissible to procreate in order to create a companion for the older child with the relatively trivial benefits above, then why would it not also be permissible to procreate in order to provide them with a benefit that is absolutely essential to their life?

It is a perhaps uncomfortable truth that parents rarely procreate purely for the sake of the child. In fact, some philosophers doubt that it is even possible to do so, because prior to conception there is no person for whose sake the parents can act.[28] If you ask modern-day Western parents why they chose to have children, they will usually cite reasons such as wanting: the personal fulfillment of parenting; to pass on genes; a playmate for existing children; to fulfill a religious

[26] Quoted in Boyle and Savulescu, "Ethics of Using Preimplantation Genetic Diagnosis to Select a Stem Cell Donor for an Existing Person," 1241.

[27] Sheldon and Wilkinson, "Should Selecting Saviour Siblings Be Banned?," 533.

[28] Mills, "Are There Morally Problematic Reasons for Having Children?"; Weinberg, *The Risk of a Lifetime*.

obligation; to satisfy grandparents; or caregivers in old age. This list of reasons is probably not exhaustive, but we think they are the most common ones.[29] What is notable is that all these reasons are instrumental. They all "use" the child to a certain extent as a means to some end, such as the happiness of siblings or to please parents, other family members, God, and so on—none of them have anything to do with the child's own interests. The concern with instrumentalization, then, does not at first sight seem particular to the savior sibling case.

Of course, the fact that even in Standard Child cases of procreation the reasons for procreating are often instrumental does not demonstrate that they are morally permissible reasons. It may be the case that it is *never* permissible to procreate for instrumental reasons and that saving an existing child's life is just one of many instrumental reasons parents should not use when deciding to have a child. If so, savior siblings are not special and we can stop here having proven the case. However, Claudia Mills argues persuasively that although it may at first seem undesirable to be valued for instrumental reasons, this is misleading.[30] And while Mills thinks the conditions that make instrumentalization acceptable in cases of procreation do not obtain in the case of savior siblings, we disagree.

Mills argues that in personal relationships we do and should desire to be valued at least in some sense instrumentally. She asks us to imagine that we are invited to dinner by a friend. When you inquire as to the reason for the invitation, if your friend replies, "I invited you for your own sake," you would likely feel somewhat offended. Did she invite me so that I could benefit from *her* amazing company? What you really want to hear is: "I invited you because I enjoy your company. You tell the best jokes and give good advice." Mills thinks it is the selfish, instrumental answer that is gratifying to us and that we actually *want* our friends to value us for certain kinds of their own selfish reasons.[31]

Imagine you are adopting a child. It is commendable to adopt the child for the child's own sake, for humanitarian reasons, perhaps. However, adopting a child is not only not undermined but is actually *enhanced* by the presence of instrumental selfish reasons. It is the difference between later saying to your child, "I adopted you to save you from a life of suffering," and, "I adopted you to save you from a life of suffering but also because I knew you would bring so

[29] In our admittedly unscientific social-media polls, no respondent gave any reason other than those listed. However, in some non-Western cultures there are likely to be additional reasons related to high infant mortality rates and a need for help on small family farms or to earn money to support the family. See also Overall, "Reasons to Have Children—Or Not," 149–51.

[30] Mills, "Are There Morally Problematic Reasons for Having Children?"

[31] Mills, "Are There Morally Problematic Reasons for Having Children?," 4.

much joy to my life." The latter is a mutually beneficial scenario: "I want you for your own sake but also for my own. We need each other."

However, it is not the case that any and all instrumental reasons are acceptable. Returning to the dinner invitation, you would not be satisfied by a response such as "I invited you for dinner to convince you to drive me to work every day." Likewise, it is not acceptable to adopt a child to be your live-in housekeeper. How do we determine which instrumental reasons are permissible and which are not?

One reason to object to the second reason for the dinner invitation and not the first is that the good that you are providing in the second case is not unique to you. Anyone could offer your friend a ride, but not just anybody could make him laugh and give him good advice. If you were to find out that all along your "friend" only continued the relationship due to interchangeable goods you provided—rides to work, help moving, and so on—you would likely feel that you were not really friends at all, because part of genuine friendship involves the *reciprocal* exchange of *noninterchangeable* goods. This points to the two criteria for an acceptable instrumental reason—it must not undermine the quality of the relationship itself by instrumentalizing the other for benefits that are nonreciprocal or goods that are extrinsic to them.

Benefits are nonreciprocal when one party receives goods from the relationship and the other does not. Having a friend help you move but then not returning the favor when they ask for help six months later is exploiting a friendship in order to receive a one-sided benefit. But even where the exchange of relationship goods is reciprocal, so not one-sided, the relationship is still undermined if one or both parties are engaged in it only for interchangeable goods. If you drove your friend to work every week (good for them) and enjoyed their company (good for you), that is reciprocal, but you would likely still feel hurt if you found out that they found your company neutral at best and continued the friendship only for the free ride to work that anyone with a car could have provided them.

Returning to procreation, if parents create a child in order to use her as a housekeeper, treating her just well enough to avoid the involvement of child protective services, they have impermissibly instrumentalized her both because the benefits of the relationship are one-sided (in the parents' favor) and the goods she provides (housekeeping services) are extrinsic to her—after all, anybody could mop the floor.

Some defenders of the practice of creating savior siblings think that while it makes the instrumental nature of reproduction more obvious, it is no less justified than other instrumentally justified procreation.[32] Is this right, or does

[32] Robertson et al., "Conception to Obtain Hematopoietic Stem Cells," 36.

creating a savior sibling constitute a special case of wrongful instrumentalization? Mills thinks it might. First, she thinks that while there are plenty of instrumental reasons for having children that create reciprocal benefits, she doubts that savior siblings create reciprocal benefits. If parents decide to have a second child so that their first child can have a companion and lifelong support, the benefit there is reciprocal because both siblings are benefiting the other in the same way. By contrast, Mills doubts that there is a reciprocal exchange of benefits in the savior sibling case because the proposed benefit (life for the sick child) is one-sided. The savior sibling is intended to be a donor for the existing child but not vice versa.[33] We are not so sure.

First, note that even if Mills is right that there is no reciprocal exchange of benefits in the savior sibling case, there is not necessarily a reciprocal exchange of benefits in Standard Child cases either. The most likely case of a reciprocal exchange of benefits is probably the Companion example. Imagine that the older child loves their sibling but the younger child strongly dislikes them, and as the two grow older they become estranged. This would not be particularly unusual; after all, you do not choose your family! In this case there is no reciprocal exchange of benefits, so perhaps it was impermissible for the parents to have the second child to provide a companion for the first. But this seems to peg the permissibility of the reproductive act on the outcome, not the parents' motivations, which were honorable—to create reciprocal benefits for both children. This seems wrong, since what nonconsequentialists find impermissible about instrumentalizing people is not what ultimately happens to them but how one relates to them—that is, one's motivations or reasons for acting, not outcomes. A nonreciprocal-exchange-of-benefits outcome can result from any of the instrumental reasons for procreation, to produce a savior sibling included. Perhaps all of these motivations for procreation are wrongful if they have the undesired result, but savior siblings are not special in this regard.

Second, it is not necessarily true that there is no potential reciprocal benefit in the savior sibling case. When the savior sibling is created, they become part of the family. They give the benefit of life to the sick child but also receive benefits in return. Many philosophers see existence itself as a benefit, but even those who do not allow that there are other benefits for the savior child, such as their sibling's companionship and their parents' love. What determines whether an exchange is reciprocal is not a direct equivalence of benefits; reciprocity need not be a tit-for-tat exchange. As Lawrence C. Becker notes, perfect returns in kind would often defeat the purpose of the reciprocity in the first place: "I don't

[33] Mills, "Are There Morally Problematic Reasons for Having Children?," 6.

want a popcorn popper; that is why I gave you mine."[34] According to Becker—and we find his account convincing—what counts is that the return is both fitting and proportionate.[35] The benefits associated with cases like Companion seem to satisfy these two conditions, but we do not see why the benefits associated with the savior sibling case cannot satisfy them too.

The expected benefits for the savior sibling can be regarded as a fitting response to the benefits received by the older child, given both are connected to the welfare of persons and their ability to be involved in loving relationships. They are a return that is of the right kind, in a way that giving the savior sibling a million dollars and then putting them up for adoption would not be.

But what about proportionality? This might be what motivates someone who finds the savior sibling case problematic, given that the act of saving a life clearly only goes one way. It might be true that no level of benefits received by the savior sibling could ever be commensurate with the benefit they give to their older sibling. But this does not mean the proportionality condition of reciprocity is violated, as often proportionality in terms of costs or effort seems just as appropriate as proportionality of benefit. For example, imagine your neighbor returns your dog that went missing. You surely are not in their debt until you rescue *their* dog. If, say, your neighbor "found" your dog simply because it walked into their yard while they were enjoying a picnic, a "thank you" seems proportionate. If, however, they found your dog after joining you on a citywide search all night, then a bigger gesture of thanks seems called for. Despite giving a great benefit to their sibling, a savior child has not put an inordinate effort (and in the cord blood case they have not put any effort) into creating those benefits. (Arguably, the effort is made by the mother through her pregnancy). So, even if a savior sibling could never receive a benefit that is strictly commensurate with the benefit they give to their older sibling, this would not exclude the possibility of a reciprocal relationship. The savior sibling case, then, is not a special case of instrumentalization from nonreciprocity or one-sided benefits.

Of course, were the parents to take the biological material from the child then shut them away for eighteen years, giving them only the minimum required for life, then the child would not enjoy any reciprocal benefits and this would be wrongful instrumentalization. But we could concoct similar forms of treatment in relation to all of the other possible ways parents decide to have children for instrumental reasons in Standard Child cases. What counts is that a child conceived for instrumental reasons (which, remember, is likely almost all children) can plausibly expect a reciprocal return of benefit, and in

[34] Becker, *Reciprocity*, 107–8.

[35] Becker, *Reciprocity*, 106–15.

the savior sibling case they can. Furthermore, as noted earlier, it borders on the absurd to think that parents who love their first child so much that they would consider going through IVF, embryo selection, and pregnancy in order to save that child's life could be the same parents who would treat the child that saved the first child so callously that that second child would receive no reciprocal benefits from their creation.

However, Mills also thinks that the savior sibling case might be a wrongful form of instrumentalization because the benefits involved are *external* to the relationship itself. Creating someone for their particular genetic profile is not instrumentally valuing them for their unique contribution to the relationship, the argument goes, but valuing them for something that anyone could provide (like valuing a friendship only for the free car rides to work). As Mills recognizes, the puzzling point here is that in deciding to procreate, even in Standard Child cases, *no* parent can value their child for themselves, given they know not a single thing about them.[36] As such, when one decides to procreate for instrumental reasons, there will always be *some* acknowledgment of the interchangeability of goods, since it is expected that any child of the set of possible children will be able to provide the benefit. Consequently, the fact that the benefit received from the donation of a particular genetic profile is interchangeable (anyone with the right profile could, in theory, provide it) does not make the savior sibling case different from other cases of procreation for instrumental reasons.

However, what Mills thinks *does* introduce a relevant difference between the savior sibling case and acceptable forms of instrumental procreation is the fact that the goods provided in the savior sibling case are not intrinsic to the parent-child relationship.[37] In Companion, the benefit received (lifelong support and love between two siblings) is something that can only be provided by having another child. But if one, say, decided to have a child only for the security of being cared for in old age, the benefit received would not be unique to the relationship (it need not be a child who takes care of you when you are elderly).

Our reply here is twofold. First, in the case considered here—parents of children with leukemia and anemia—savior siblings often *are* the only persons who can provide such a benefit. If alternative donors were available, the savior sibling would not be needed. As such, the benefit *is* unique to the parent-child relationship. Second, the motivations for choosing to have a savior sibling *themselves* originate from a parent-child relationship. It is out of a concern to prevent their other child from dying that the parents decide to have a savior sibling. And so, while the child might be created for instrumental reasons, the child is not

[36] Mills, "Are There Morally Problematic Reasons for Having Children?," 8.

[37] Mills, "Are There Morally Problematic Reasons for Having Children?," 9.

created for reasons inimical or opposed to valuable relations between a parent and their child—it is because the parent loves their child unconditionally that they decide to have another child. Of course, if the parents put the savior sibling up for adoption immediately following donation, that would express that the savior sibling was only valued instrumentally for something external to an appropriate parent-child relationship (never mind that that is an unacceptable way to treat a child). But again, the savior sibling case is not special in this regard. We would say the same thing if, in Companion, the child were put up for adoption once it became obvious the sibling relationship was not working out.

These comments show that savior siblings do not present a special form of instrumentalization. First, they can expect a reciprocal return of benefits. Second, because there are no other options for donation available and the motivation for the savior sibling's creation is internal to a parental-child relationship, they are not valued merely for providing benefits that are interchangeable. We think what is really driving the instrumentalization objection is a worry that the child will be mistreated somehow, that parents will not relate to their child in the right way. But again, rather than being an argument that savior siblings *per se* present a special case, this worry indicates that the important factor is not the reasons for the child's conception but the child's treatment after birth. No one worries about the children created to be companions, nor about the children created for the array of other instrumental reasons that make up the Standard Child case. Why do we worry about savior siblings?[38]

3. CONCLUSIONS

There is clearly something different about savior siblings compared to other children—they are created to save another. Our claim in this paper has been that this descriptive difference does not raise special normative issues of procreative permissibility and impermissibility. On the contrary, the conditions in which it is permissible or impermissible to create a savior sibling are the same conditions in which it is permissible or impermissible to create any child.

[38] One might argue that there is something else special about savior siblings—viz., that they are genetically selected for the benefit of a third party. For this objection to hold, it must be the case that savior siblings are special compared to the Standard Child just in case they are genetically selected to help a third party in a way that a Standard Child is not. We have already argued that creating a child to benefit a third party is not necessarily impermissible, so it must be the fact that the child was genetically selected that is relevant. This takes us into a more general bioethical debate about the permissibility of genetic selection itself. If genetic selection for anything other than disease prevention is impermissible, then savior siblings likely are too, but not because they are savior siblings. Thank you to an anonymous reviewer for suggesting this objection.

We first argued that either there is nothing inherent in the creation of savior siblings that will lead to physical harm (the cord blood case), or, if there were physical harm (the bone marrow or organ donation case), the procedure would only ever be permissible if the same ethical and legal requirements that exist in Standard Child cases were met. Furthermore, given that bone marrow and organ donation are only possible and not inevitable outcomes, the savior sibling case is no different from all those instances of the Standard Child case where parents choose to create a child in the knowledge that the child may be harmed and/or experience pain in the future. We then argued that savior siblings are no different from Standard Child cases of procreation when it comes to the possibility they will experience psychological harm, whether that results from their knowledge of the reason for their existence or from the treatment they receive from their parents.

We also argued that a child's autonomy is no more undermined in the savior sibling case than it is in the Standard Child case. It is, of course, true that savior siblings consent neither to being born for the purpose of donation nor to undergoing medical treatment at a stage of precompetence. But this does not make savior siblings special, because no child consents to be born for any purpose, and parents make decisions about their precompetent children's medical treatment all the time. Furthermore, while it might be thought that being created for the purpose of blood donation undermines the child's right to an "open future" by making them more likely to make further donations later in their life, or that any choice to donate an organ by a savior sibling can never be a case of "informed consent," this thought results either from an erroneous conflation of increased likelihood to donate with involuntariness or from unacceptable pressuring from parents, the latter of which also violates the autonomy of children in Standard Child cases.

As we outlined, the reasons driving most (all?) procreation are instrumental. And this instrumentalization need not make reasons for procreation wrongful but can actually be a good thing, so long as the benefits involved are reciprocal and noninterchangeable. And we argued that there is nothing stopping the instrumental reasons underpinning the savior sibling case from meeting these two conditions. What seems to be driving concerns about instrumentalizing savior siblings is a worry that the child will be mistreated somehow. A child being mistreated would, of course, be terrible, but it would be terrible because the child was mistreated, not because they were instrumentalized. In fact, if it were possible to create children for their own benefit (though, as we said, many doubt this), it would be no less terrible for the child who is born for this reason to be mistreated than for a child created for some acceptable instrumental purpose (such as a savior sibling) to be mistreated.

What the argument highlights is that any wrong-making features in the creation of savior siblings are no different from those in Standard Child cases, in particular, the quality of the child's life (including how they are treated and related to) once they exist. In other words, what would make having created a savior sibling wrong would be no different from what would make the creation of a Standard Child wrong—neglect, abuse, lack of love, lack of respect for the child's autonomy, and so on.[39]

However, although we have argued that the unique reason for their birth does not affect the moral permissibility of creating them, this special feature of savior siblings might have normative implications for whether their creation may actually be morally obligatory. In the case considered here (where no other donors are available), creating a savior sibling is the only option for preventing a significant bad from happening (a child dying). As such, answering whether the practice is morally obligatory will need to be sensitive to how we weigh the prevention of such a bad against the costs imposed by the practice on parents (especially mothers). These are clearly both morally weighty reasons, and it is not immediately obvious how such a weighting would best be made. Unfortunately, examining the implications of this special feature of savior siblings must be a task left for future work.[40]

Trinity College Dublin
althorpc@tcd.ie

Western University
efinnero@uwo.ca

REFERENCES

Beauchamp, Tom L., and James F. Childress. *Principles of Biomedical Ethics*. 6th ed. Oxford: Oxford University Press, 2008.

Becker, Lawrence C. *Reciprocity*. Chicago: University of Chicago Press, 1990.

[39] Spelling out exactly what children are entitled to expect from their parents on pains of having been wronged is beyond the scope of this paper, but promising accounts can be found in S. Matthew Liao's *The Right to Be Loved* and Erik Magnusson's "Children's Rights and the Non-Identity Problem."

[40] The authors contributed equally to this paper. An earlier version was presented at the 2022 Bucharest Conference in Analytical Political Theory, and we would like to thank participants for their helpful questions and discussion. We would also like to thank the two external reviewers for their detailed comments, which led to an improved paper.

Benatar, David. *Better Never to Have Been: The Harm of Coming into Existence*. Oxford: Oxford University Press, 2006.

Bizzetto, Renata, Carmen Bonfim, Vanderson Rocha, Gérard Socié, Franco Locatelli, KaWah Chan, Oscar Ramirez, et al. "Outcomes after Related and Unrelated Umbilical Cord Blood Transplantation for Hereditary Bone Marrow Failure Syndromes Other Than Fanconi Anemia." *Haematologica* 96, no. 1 (January 2011): 134–41.

Boyle, Robert J., and Julian Savulescu. "Ethics of Using Preimplantation Genetic Diagnosis to Select a Stem Cell Donor for an Existing Person." *British Medical Journal* 323, no. 7323 (November 2001): 1240–43.

Brierley, Joe, and Vic Larcher. "Organ Donation from Children: Time for Legal, Ethical and Cultural Change." *Acta Paediatrica* 100, no. 9 (September 2011): 1175–79.

Campbell, M., L. Wright, R. A. Greenberg, and D. Grant. "How Young Is Too Young to Be a Living Donor?" *American Journal of Transplantation* 13, no. 7 (July 2013): 1643–49.

Chan, Tak Kwong, and George Lim Tipoe. "The Policy Statement of the American Academy of Pediatrics—Children as Hematopoietic Stem Cell Donors—A Proposal of Modifications for Application in the UK." *BMC Medical Ethics* 14, no. 1 (October 2013).

Clayton, Matthew. "The Case against the Comprehensive Enrolment of Children." *The Journal of Political Philosophy* 20, no. 3 (September 2012): 353–64.

Committee on Bioethics. "Children as Hematopoietic Stem Cell Donors." *Pediatrics* 125, no. 2 (February 2010): 392–404.

Dillon, Robin S. "Self-Respect and Self-Esteem." In *International Encyclopaedia of Ethics*, edited by Hugh LaFollette. Hoboken, NJ: John Wiley & Sons, 2019. https://doi.org/10.1002/9781444367072.wbiee221.pub2.

Dworkin, Gerald. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press, 1988.

Dworkin, Ronald. "The Original Position." In *Reading Rawls: Critical Studies on Rawls' "A Theory of Justice,"* edited by Norman Daniels, 16–53. New York: Basic Books, 1975.

Enoch, David. "Hypothetical Consent and the Value(s) of Autonomy." *Ethics* 128, no. 1 (October 2017): 6–36.

Eyal, Nir. "Informed Consent." In *Stanford Encyclopedia of Philosophy* (Spring 2019). https://plato.stanford.edu/archives/spr2019/entries/informed-consent/.

Feinberg, Joel. "The Child's Right to an Open Future." In *Freedom and Fulfillment: Philosophical Essays*, 76–97. Princeton: Princeton University Press, 1992.

Hannan, Sarah. “Childhood and Autonomy.” In *The Routledge Handbook of the Philosophy of Childhood and Children*, edited by Anca Gheaus, Gideon Calder, and Jurgen De Wispelaere, 112–22. New York: Routledge, 2018.

Haude, K., P. McCarthy Veach, B. Leroy, and H. Zierhut. “Factors Influencing the Decision-Making Process and Long-Term Interpersonal Outcomes for Parents Who Undergo Preimplantation Genetic Diagnosis for Fanconi Anemia: A Qualitative Investigation.” *Journal of Genetic Counseling* 26, no. 3 (June 2017): 640–55.

Kakourou, Georgia, Christina Vrettou, Maria Moutafi, and Joanne Traeger-Synodinos. “Pre-Implantation HLA Matching: The Production of a Saviour Child.” *Best Practice & Research Clinical Obstetrics and Gynaecology* 44 (October 2017): 76–89.

Leung, Wing, Dario Campana, Jie Yang, et al. “High Success Rate of Hematopoietic Cell Transplantation Regardless of Donor Source in Children with Very High-Risk Leukemia.” *Blood* 118, no. 2 (July 2011): 223–30.

Liao, S. Matthew. *The Right to Be Loved*. New York: Oxford University Press, 2015.

Locatelli, Franco. “Outcome of Patients with Hemoglobinopathies Given either Cord Blood or Bone Marrow Transplantation from an HLA-Identical Sibling.” *Blood* 122, no. 6 (August 2013): 1072–78.

Magnusson, Erik. “Children’s Rights and the Non-Identity Problem.” *Canadian Journal of Philosophy* 49, no. 5 (August 2019): 580–605.

Mills, Claudia. “Are There Morally Problematic Reasons for Having Children?” *Philosophy and Public Policy Quarterly* 25, no. 4 (Fall 2005): 2–9.

Mullin, Amy. “Children, Autonomy, and Care.” *Journal of Social Philosophy* 38, no. 4 (Winter 2007): 536–53.

Overall, Christine. “Reasons to Have Children—Or Not.” In *The Routledge Handbook of the Philosophy of Childhood and Children*, edited by Anca Gheaus, Gideon Calder, and Jurgen De Wispelaere, 147–57. London: Routledge, 2018.

Richards, Janet Radcliffe. “A World of Transferable Parts.” In *A Companion to Bioethics*. 2nd ed., edited by Helga Kuhse and Peter Singer, 375–89. Chichester: Wiley-Blackwell, 2009.

Robertson, John A., Jeffrey P. Kahn, and John E. Wagner. “Conception to Obtain Hematopoietic Stem Cells.” *Hastings Center Report* 32, no. 3 (May–June 2002): 34–40.

Rocha, Vanderson, John E. Wagner, Kathleen A. Sobocinski, and John P. Klein. “Graft-Versus-Host Disease in Children Who Have Received a Cord-Blood or Bone Marrow Transplant from an HLA-Identical Sibling.” *The New England Journal of Medicine* 342, no. 25 (June 2000): 1846–54.

Rubeis, Giovanni, and Florian Steger. "Saving Whom? The Ethical Challenges of Harvesting Tissue from Savior Siblings." *European Journal of Haematology* 103, no. 5 (November 2019): 478–82.

Sachs, David. "How to Distinguish Self-Respect from Self-Esteem." *Philosophy and Public Affairs* 10, no. 4 (Fall 1981): 346–60.

Saunders, Ben. "Consent and Organ Donation." In *The Routledge Handbook of the Ethics of Consent*, edited by Peter Schaber and Andreas Müller, 311–21. London: Routledge, 2018.

Sheldon, Sally, and Sheldon Wilkinson. "Should Selecting Saviour Siblings Be Banned?" *Journal of Medical Ethics* 30, no. 6 (December 2004): 533–37.

Strong, Kimberly A., Chris F. Jordens, Ian H. Kerridge, John Miles Little, and Rachel A. Ankeny. "It's Time to Reframe the Savior Sibling Debate." *AJOB Primary Research* 2, no. 3 (October 2011): 13–25.

Weinberg, Rivka. *The Risk of a Lifetime: How, When, and Why Procreation May Be Permissible*. Oxford: Oxford University Press, 2016.

Wilkinson, T. M. *Ethics and the Acquisition of Organs*. Oxford: Oxford University Press, 2011.

Wolf, Susan M., Jeffrey P. Kahn, and John E. Wagner. "Using Preimplantation Genetic Diagnosis to Create a Stem Cell Donor: Issues, Guidelines and Limits." *Journal of Law, Medicine, and Ethics* 31, no. 3 (Fall 2003): 327–39.

Journal of Ethics and Social Philosophy
vol. 26, no. 1 · november 2023

https://doi.org/10.26556/jesp.v26i1.2513

THE ETHICS OF CONTINUING HARM

Joseph Chapa

"I'm a kidnapper for her, that's what I am," Tumnus tells Lucy. Shortly after Lucy Pevensie arrives in Narnia, she is befriended—or so she thinks—by a Faun named Mr. Tumnus. Tumnus lures her into his home and delivers on his promises of tea and a warm fire. Lucy, unaware, is not his guest but his prisoner. Tumnus's acts of apparent hospitality are in fact stalling tactics as he awaits the arrival of the White Witch who will undoubtedly kill Lucy. In a moment of contrition and on the verge of releasing his prisoner, Tumnus confesses. He is a kidnapper.

> "Well," said Lucy rather slowly, . . . "that was pretty bad. But you're so sorry for it that I'm sure you will never do it again."
>
> "Daughter of Eve, don't you understand?" said the Faun. "It isn't something I have done. I'm doing it now, this very moment."[1]

Tumnus's confession presents a puzzle. What can he mean when he says that he is now, at this very moment, in the act of kidnapping? Was he in the act of kidnapping when he built the fire and put on the kettle? Was he in the act of kidnapping while he played his flute and she slept? In this moment, even as Tumnus confesses to Lucy, she is nevertheless his prisoner. Tumnus continually violates Lucy's right not to be kidnapped—the harm he poses is a continuing harm. Moreover, in this moment, Tumnus also poses a continuing threat of harm, for, at each moment, he is about to imprison her even longer and thereby to harm her even further. The harm Tumnus causes and the threat he poses are temporally coextensive—his is a continuing harm and a continuing threat. He is correct both about the harm and the threat when he says he is "doing it now, this very moment." And, crucially, the continuing violation of Lucy's right to freedom he has caused is in addition to the lethal harm the Witch would have caused. Throughout her imprisonment, Tumnus continually violated Lucy's rights even though he never turned her over to be killed.

[1] Lewis, *The Lion, the Witch and the Wardrobe*, 16.

At first glance, the question of defensive harming might appear to be merely a question of whether an imminent threat is a necessary condition for justified defensive harming. For instance, one might think that the threat Tumnus poses is fully explained as an ordinary nonimminent threat, and that nonimminent threats can justify defensive harming. Instead, as I argue here, the continuing threat attackers can pose, and the continual harm attackers can cause, are of a different kind than the discrete threats we often think of as either imminent or nonimminent. The harm victims suffer, and therefore the harm defenders may proportionately cause to defeat those threats, are fully captured only if we include continuing harms.

In the literature on the ethics of defensive harming, threats are often categorized as either imminent or nonimminent; and nonimminent threats are often treated as though the harms they threaten are discrete, rather than continuous. In other words, the existing literature on defensive harming sometimes does include threats of continuing harm, but those threats of continuing harm are often coincident with threats of discrete harm. As a result, it is difficult to determine what role threats of continuing harm play in moral justifications for defensive harming. As I argue here, a theory of defensive harming is incomplete if it fails also to identify and account for threats of continuing harm like the one Tumnus posed to Lucy. To this end, I develop continuing harm as a subset of the harms that are relevant to defensive harming. I then sketch an account of the morality of continuing harms and show that the proportionality calculus in defensive harming is sensitive to threats of continuing harm in addition to threats that are more easily categorized as imminent or nonimminent. I pay special attention at the end of the paper to the application of this account of continuing harm to the proportionality condition in just war thinking. To summarize, I intend for this paper to serve two purposes: first to show that we should develop an account of continuing harms, and then to develop one such account.

There are several ways in which harms caused over time can be relevant to questions of proportionality. For instance, in the context of war, one might ask whether, or in what ways, harms that have already been caused are relevant to a forward-looking proportionality calculus. That is, suppose a state decides that it can proportionately accept one thousand friendly losses to achieve its just war aims. Further suppose that once the fighting begins, the state has suffered nine hundred losses. It might still achieve its aims, but doing so will likely result in five hundred more losses. Is the state morally justified in continuing its war?[2] Though these questions certainly involve proportionality and harms caused

[2] These are questions Fabre, McMahan, and Rodin have addressed in a 2015 *Ethics* symposium on ending wars. See Fabre, "War Exit"; McMahan, "Proportionality and Time"; Rodin, "The War Trap."

over time, they are peripheral to the questions of continuing harm with which I am concerned here. One difficulty in discussing continuing threats and continuing harms is that these terms can be ambiguous. For instance, an attacker might continuously threaten a discrete harm. Suppose an attacker says, "As soon as you're alone, I'm going to break your arm." If the attacker follows through on the threat, the harm he causes is discrete; and yet as the attacker lurks, waiting for his victim to be alone, the threat is continuous. Or suppose an attacker unjustly and imminently threatens to amputate my arm and follows through. I will suffer some immediate harm—pain, for example—but the deprivation of the use of my arm will also amount to a continuing harm for the rest of my life. Here, the threat is discrete, but the harm is continuing. Or an attacker might pose a discrete threat of a continuing harm. In Singer's torture case, an attacker sets the torture machine in motion, continually harming his victim, but then dies.[3] The harm the victim suffers is continuous even when the threat ceases to be continuous. Throughout this paper, when I refer to "threats of continuing harm," I have in mind situations in which an attacker threatens to deprive his victim of some right continually, for instance, a right to freedom. Often, though not always, when an attacker causes a continuing unjust harm, the attacker also poses a continuing threat. In this paper, I focus on threats of continuing harm as opposed to continuing threats of harm even though the latter often accompany the former. According to this definition, kidnapping victims, hostages, and those unjustly imprisoned are all victims of continuing threats.[4]

These kinds of continuing harms—kidnapping, hostage-taking, and imprisonment—are different from repeated discrete harms. For instance, imagine a victim who is unjustly imprisoned by an aggressor. Each day of the imprisonment, the aggressor beats the victim. There are (at least) two different kinds of unjust harms being perpetrated against the victim. First, she is regularly and repeatedly being beaten. These are repeated discrete harms. But additionally, she is unjustly imprisoned. In most cases—both hypothetical and real-world cases—continuing harms are often accompanied by repeated discrete harms. Though a discussion of repeated discrete harms will certainly arise throughout the paper, I am focused on the continuing harms of, for instance, kidnapping, hostage-taking, and imprisonment.

Ultimately, as I argue below, continuing harms are relevant to the proportionality calculus in defensive harming and in just war thinking. If we fail to account for continuing harms, we also fail to account for the role that the

[3] Singer, "Bystanders to Poverty," 195.

[4] There are also justified threats of continuing harm. If someone is justifiably placed in prison, for example, she faces a justified threat of continuing harm. If she kills to defeat that threat, she does so impermissibly.

duration of the relevant harm can play in the proportionality calculus. One is undoubtedly permitted to cause more harm to keep from being wrongfully imprisoned for a year than one is permitted to cause to keep from being wrongfully imprisoned for a day. To explain this difference, we must appeal not just to repeated threats of discrete harm that can be easily categorized as threats of imminent and nonimminent harm, but specifically to threats of continuing harm. Common approaches to defensive harming and, more specifically, to just war theory, have not thoroughly developed threats of continuing harm.

1. Continuing Harms in the Real World

The fact that threats of continuing harm obtain in the real world—and that legal categories might be insufficient to account for them—arises in several real-world cases of domestic abuse. Here, I focus on Judy Norman's case.

> *Domestic Abuse*: Judy Norman was tortured and beaten by her husband for years and threatened with death and mutilation if she attempted to escape or obtain outside help. She finally killed her sleeping husband. Denied a self-defense instruction [by the court], she was convicted of manslaughter and sentenced to six years' imprisonment.[5]

There is a widely held intuition that Judy Norman was justified in killing her abuser, but it is not immediately obvious what moral principle ought to ground that justification.

On the night that Norman killed her abuser, she did not face an imminent threat. Throughout the paper, by "imminent threat" I mean that an attacker will harm the victim immediately. The time available for a defender to react is so brief that her defensive options are drastically reduced. In Kimberly Kessler-Ferzan's words, "an imminent threat is one that will happen 'in an instant' or 'at once.'"[6] The intuition that Norman was justified in killing her abuser stands in contrast to the common legal requirement that either a justified defensive harm must be immediately necessary or that defensive harm is justified only against imminent threats. This conflict between moral intuition and legal standards is reflected in the conflicting legal decisions on the case. The district court initially denied a self-defense instruction to the jury, and Norman was convicted. The appellate court, however, reversed the decision arguing that the district court

[5] Rosen, "On Self-Defense, Imminence, and Women Who Kill Their Batterers," 372. Norman's abuse is cataloged in greater detail in Fletcher, "Domination in the Theory of Justification and Excuse," 555; and Ferzan, "Defending Imminence: From Battered Women to Iraq," 233–34.

[6] Ferzan, "Defending Imminence," 229.

erred and that the jury should have been given a self-defense instruction. Later, the Supreme Court of North Carolina reversed again, arguing that the self-defense instruction was not available for Norman precisely because the imminence requirement had not been met.[7] One plausible—if partial—explanation for these conflicting legal interpretations is that imminent threats are not the only threats relevant to moral justifications for defensive harming.

Though Norman did not face an imminent threat on the night she killed her abuser, she almost certainly faced a nonimminent threat of discrete harm. Norman's husband had already threatened her "with death and mutilation if she attempted to escape or obtain outside help."[8] Suppose that, on the night she killed her abuser, he credibly told her that he would beat her the next morning. If so, she faces a threat of nonimminent discrete harm. Perhaps it is this threat that justifies her killing her abuser. On this question, too, various legal jurisdictions disagree with one another. For example, the English legal system retains imminent threat as a necessary condition for a self-defense justification but has considered some cases of victims killing their abusers in nonconfrontational situations as cases of provocation, and therefore treats abuse victims who kill their abusers more leniently. Australia, by contrast, has abandoned imminent threat as a necessary condition in cases in which the abuse victim who kills is diagnosed with "battered woman syndrome." Finally, the various jurisdictions in the United States have adopted different positions. Some retain imminent threat as a necessary condition for justified defensive harming and others have rejected it. George Fletcher went so far as to say as early as 1996 that "the central debate in the theory of self-defense for the last decade has been whether we should maintain a strict requirement of imminence in assessing which attacks trigger a legitimate defensive response."[9]

Even if we agree, though, that Norman faced no imminent threat but did face a nonimminent threat of discrete harms of physical assault, our understanding of Norman's moral justification is incomplete without incorporating the threat of continuing harm she faced. Norman was not the victim solely of physical torture, but she was also, in Fletcher's words, made to be a prisoner in "this gulag she called home."[10] If the intuition that Norman was morally justified in killing her abuser is correct, and she faced no imminent threat but instead faced nonimminent threats of discrete harm and a threat of continuing harm,

[7] Ferzan, "Defending Imminence," 235.

[8] Rosen, "On Self-Defense, Imminence, and Women Who Kill Their Batterers," 372.

[9] See, respectively, Belew, "Killing One's Abuser," 770, 787–88; Fletcher, "Justification and Excuse," 567.

[10] Fletcher, "Justification and Excuse," 556.

it is unclear whether either the nonimminent threats or the continuing threats are independently sufficient to justify defensive harming. As I argue here, Norman's justification for defensively killing her husband instead is grounded, at least in part, in the threat of continuing harm—what Ferzan calls "the kidnapping paradigm."[11] At the moment Norman killed her husband, he posed a threat of nonimminent harm in that he had previously threatened to kill and mutilate her; but he also posed a threat of continuing harm in that he made her—even at that moment—a prisoner in her own home.

Some might be skeptical of the claim that Norman was made to be a prisoner in her own home. After all, she was not physically restrained nor locked up. While her husband slept, it was physically possible for her to walk out the front door—even if doing so would have resulted in additional beatings. This had, in fact, occurred. In Fletcher's words, "she had experienced beatings in retaliation for prior efforts to leave the scene of her suffering."[12] And so, if she is physically able to leave the house, even if doing so would result in additional harm, then is she really a prisoner? Or is it more accurate to describe her condition only with reference to the several discrete harms that would result if she tried to leave the house?[13] This is an important question because, if the whole of her suffering can be captured with reference to discrete harms, then the concept of continuing harm adds no additional value. But I do not think that the fact that she can walk away, at the risk of additional discrete harms, entails that she is not imprisoned, and therefore, it does not entail that she suffers no continuing harm. Instead, the conditional threat that escape attempts will result in additional harms is one element of her imprisonment. Indeed, threats of additional harm in response to escape attempts are probably a common element of imprisonment. In the US federal detention system, for example, "escape" is considered to be a prohibited act of the "greatest severity," and sanctions for escapees can include anything from monetary fines to extended imprisonment.[14] Like Norman, if a prisoner in the federal penal system for whom it is physically possible to escape does walk away, he risks suffering additional discrete harms as a result. The fact that a victim might possibly—or even might easily—escape does not entail that the victim is not imprisoned. Had Norman attempted to escape, it is very likely that she would have been punished. We should not conclude from this that she was not imprisoned; only that punishment for attempting escape was a feature of her imprisonment.

[11] Ferzan, "Defending Imminence," 253.

[12] Fletcher, "Justification and Excuse," 555.

[13] I am grateful to an anonymous reviewer for identifying this concern.

[14] 73 Fed. Reg. 76263 (Dec. 8, 2010), reprinted as amended in 28 C.F.R 541.3.

In what remains of this paper, I develop this conception of threats of continuing harm and attempt to demonstrate its relevance to the morality of defensive harming, and ultimately, to just war thinking.

2. Developing Threats of Continuing Harm

2.1. Continuing Harms in the Defensive Harming Literature

Even if we agree that Judy Norman had a moral justification for defensively killing her husband, that justification might be overdetermined. For instance, the repeated, discrete threats of torture might be sufficient on their own to justify defensive killing without reference to the continuing harm of imprisonment. Or else, perhaps the threat of continuing harm is sufficient to justify defensive killing without reference to the several threats of discrete harm. One reason it is difficult to develop an account of continuing harms is that in most cases, both real and hypothetical, continuing harms are accompanied by discrete harms. In such cases, it is impossible to determine whether the nonimminent threat of discrete harm or the threat of continuing harm is independently sufficient to justify defensive harm. Threats of continuing harm do, indeed, arise in the defensive harming literature, but when they do, they are accompanied by threats of discrete harm. Consider these two well-known cases.

Here is Helen Frowe's Trolleyology case.

> *Trolleyology*: Imagine that I . . . lock you up in your house for an indefinite period, perhaps the rest of your life. I force you to practice a particular religion. . . . I make you dress in army camouflage, read only the collected works of Frances Kamm, spend hours enacting moral dilemmas using a toy train set, and start every day with a rousing rendition of what I like to call the Trolley Anthem. I credibly threaten to kill you if you try to force me out of the house or otherwise resist the imposition of my regime.[15]

Frowe's own intuition in Trolleyology is that her victim is indeed permitted to kill to defeat the threat she asks us to imagine. But which of the myriad hypothetical harms Frowe imposes on her victim are sufficient to justify defensive killing? Is the victim justified in defense of his right to autonomy each discrete time Frowe threatens his autonomy? Or is he justified because she poses a conditional threat of nonimminent lethal harm if he resists? Or is he justified on the grounds that she has locked him up in the first instance, thereby continually violating his right to personal freedom? Or is it that the threats are independently insufficient to justify lethal defensive harm, but jointly sufficient?

[15] Frowe, *Defensive Killing*, 140–41.

In other words, it is unclear whether the threat of continuing harm by itself is doing the work in the intuitive response to the case.

Fabre's Home Invaders case avoids the additional threat of imminent harm but still includes a conditional threat of nonimminent harm alongside the threat of continuing harm.

> *Home Invaders*: Suppose that your house is wrongfully invaded by a group of individuals who intend to stay there permanently and who use coercive force against you if you dissent with whatever decision they make with respect to your house. . . . [And suppose that the home invaders coerce you into living a less than minimally decent life.] You have a choice between fully complying with wrongdoers' demands or killing them.[16]

There are at least two kinds of harm at issue here. The victim is imprisoned, as in Trollyology. But additionally, the victim is coerced into living a "less than minimally decent life." Though we do not know exactly what harms constitute or have brought about that condition, whatever they are, they are in addition to the continuing harm of imprisonment. In other words, I believe it is possible for prisoners—say, persons rightly found guilty of white-collar crimes serving prison sentences in minimum security prisons—to live a minimally decent life. If so, then imprisonment by itself does not imply or entail a less than minimally decent life. If this is correct, then the victim in Home Invaders is the victim both of imprisonment and of other harms significant enough to make his life less than minimally decent. I agree with Fabre's conclusion that the victim has a right defensively to kill the invaders, provided the home occupation will continue for some significant period of time; but because the victim faces two kinds of threat—the continuing harm of imprisonment and whatever harms constitute the cause of his less than minimally decent life—it is not clear whether the continuing harm of imprisonment on its own is sufficient to justify defensive killing.

In each of these cases, and in Judy Norman's case, we are left wondering which, if any, of the harms imposed is sufficient to justify defensive killing. In each case, a threat of continuing harm is posed, but without developing the category of threats of continuing harm, we can have no way of knowing whether such a threat is sufficient to justify lethal defensive harm.

2.2. Continuing Threats and Proportionality

To this point, I have considered the distinction between threats of continuing harm and threats of discrete harm. But what is the relationship between threats

[16] Fabre, *Cosmopolitan War*, 69.

of continuing harm and proportionality? Threats of continuing harm can be, *inter alia*, threats of unjust imprisonment, kidnapping, or violations of personal freedom. While rights to personal freedom are important and violations of those rights significant, it nevertheless seems as though one would be permitted to cause more harm to defend one's life than one would be permitted to defend one's freedom. And if so, it is *pro tanto* impermissible to cause lethal harm to defeat a threat of continuing harm. But is this correct? A close look at a second distinction—that between vital and lesser interests—will help to answer that question.

The distinction between vital and lesser interests was first introduced by David Rodin. A vital interest is an interest the defense of which can justify lethal harm. According to Rodin,

> [Vital interests are] those centrally important interests, the unjust threat to which can justify lethal force in [the] context of self-defense. These are in broad terms: threat to life, substantial threat to bodily integrity (including loss of limb, torture, and rape), profound attacks on liberty such as slavery, and permanent or long-standing displacement from one's home.[17]

At its core, the distinction between vital and lesser interests is a descriptive mechanism for proportionality in defensive harming cases. Whether an interest falls into the lesser or the vital category depends upon what one may do to defend it. An interest is vital only if one can be, under some circumstances, justified in killing to defend it.

To determine whether threats of continuing harm are sufficient to justify lethal defensive harm—that is, whether a threat of continuing harm can threaten a vital interest—we are in need of a case in which the attacker threatens the victim with continuing harm only and not with other discrete harms.

> *Pleasant Detention*: Agatha wants wrongfully to abduct and imprison Violet. Agatha gently takes Violet prisoner in Violet's sleep and deposits her into the detention facility, cut off from the outside world. Within the detention facility, however, she is well fed, warm, and generally well provided for. She is free to do more or less as she pleases. Agatha, for unknown reasons, has rigged the locks to a pulse monitor on Agatha's own wrist. If she dies, the locks will open. Violet, by mere happenstance, finds a gun in the detention facility and is able to shoot and kill Agatha in the adjacent building in which Agatha lives whenever she chooses.

[17] Rodin, "The Myth of National Self-Defence," 80.

During Violet's imprisonment, the harm with which Agatha threatens Violet is neither like a paradigmatic case of a threat of imminent harm nor is it like a paradigmatic case of a threat of future harm. In an important sense, continuing harms are constituted by both a threat of imminent harm and a threat of future harm. In Pleasant Detention, Agatha threatens to harm Violet "in an instant" or "at once."[18] But it is also the case that, to the degree that Agatha intends to keep Violet imprisoned for some time, Agatha also threatens to harm Violet in the future. A threat of imminent harm and a threat of future harm are both constituent parts of a threat of continuing harm. And because this harm takes place over time, the defensive harm that Violet may proportionately cause to Agatha depends upon the duration of her captivity. To be denied one's freedom for an hour or a day might not justify lethal defensive harm. But cutting Violet off from the rest of the world for the rest of her life seems to reach the proportionality threshold for lethal defensive harm. Defensive killing, at least in cases like Pleasant Detention, can be permissible against threats of continuing violation of interests such as personal freedom if the duration is sufficiently lengthy.

There is an important implication here regarding the common use of "vital interests" and "lesser interests." If an interest is threatened with continuing harm, then whether that interest is vital or lesser can depend upon the duration of the threatened harm. For example, it is not sufficient to say that one's interest in not being killed is a vital interest while one's interest in not being imprisoned is a lesser interest. For, as is the case in Pleasant Detention, whether or not one may kill to defend one's interest in not being imprisoned depends in part upon the duration of the imprisonment. And since the category of vital interests is defined as the set of interests one may kill to defend, then whether one's interest in not being imprisoned is a vital interest depends on the duration of the imprisonment. Those who offer lists of exemplar vital interests often include interests in not being enslaved or kidnapped. Rodin includes in the set of vital interests "profound attacks on liberty such as slavery, and permanent or long-standing displacement from one's home." Likewise, in summarizing Rodin's position, Frowe includes slavery in her list. Lazar includes one's interest in not being kidnapped. Fabre argues that one may kill to defeat threats of kidnapping and enslavement. McMahan holds that in some cases one may kill to prevent "enslavement or captivity for a significant or indefinite period."[19]

[18] Ferzan, "Defending Imminence," 229.

[19] See, respectively, Rodin, "The Myth of National Self-Defence," 80; Frowe, *Defensive Killing*, 125; Lazar, "National Defence, Self-Defence, and the Problem of Political Aggression," 15; Fabre, *Cosmopolitan War*, 69; McMahan, "War as Self-Defense," 78.

The reason so many theorists hold that it can be proportionate to kill in defense against kidnapping and enslavement points to the moral significance of the continuing harm. One's interest in not being kidnapped or enslaved is a vital interest only if the harm that is threatened in these cases is a continuing harm of a sufficient duration. Though we might maintain the distinction between vital and lesser interests, there is no such distinction between vital and lesser rights. This is because, while the categories of vital and lesser interests are sensitive to duration, the category of rights is not. My right not to be unjustly killed is of greater moral significance than my right not to be unjustly imprisoned, and this claim is true independent of the duration of my imprisonment. But I have a much stronger interest—indeed I have a vital interest—not to be imprisoned for a decade as compared to my lesser interest not to be imprisoned for a day.

If proportionality is sensitive to the duration of a continuing harm, it should be no surprise that proportionality is also sensitive to the magnitude of the harm threatened. By "magnitude," I mean the amount of harm one suffers at any given moment. Suppose Agatha is a supervillain who can cause headaches in her victims. Causing a severe migraine for an hour is worse than causing a mild headache for an hour. The proportionality calculi track this difference. A defender would be justified in causing more harm to defend the migraine victim than one would be justified in causing to defend the mild headache victim.

The moral weight of the continuing harm is sensitive both to the magnitude of harm and the duration of harm. For proportionality, then, what is at issue is something like the "area under the curve" in integral calculus. A harm of lesser magnitude can justify greater defensive harm if continuous over a long period of time, while a harm of greater magnitude over a shorter period of time can justify that same defensive harm.

Though the moral weight of continuing harm such as imprisonment increases over time, the increase is not necessarily linear. That is, any period during which the victim suffers continuing harm might be of a different moral significance than other periods of the same duration. For instance, it might be the case that the longer one is imprisoned, the more harmful is each day of imprisonment. Or perhaps imprisoning a victim during a period that causes her to miss important events in the lives of loved ones might be worse than imprisoning a victim during a period of the same duration that does not cause her to miss important events. Or perhaps, for psychological reasons, the first few days of imprisonment are among the most difficult as the prisoner becomes accustomed to her new environment. Or maybe, as in Judy Norman's case, the discrete harms—torture, physical violence, and the like—that accompany the imprisonment make the continuing harm of imprisonment itself more harmful than it would otherwise be. My argument about the relationship between continuing harms and proportionality does not depend on answers to these nuanced questions about the causal relationships between harms. Instead, my argument is compatible with any number of potential accounts of the causal relationships between harms.

Having developed this account of continuing harms, we are in a position, first, to return to Judy Norman's case to determine what relevance continuing harms might have to her self-defense justification and, second, to apply this conception to political conflict.

2.3. *Applying Threats of Continuing Harm to Domestic Abuse Cases*

In Norman's Domestic Abuse case, as in the aforementioned Trolleyology and Home Invaders cases, it is difficult to isolate the various kinds of threat and harm to determine which, if any, is independently sufficient to justify lethal defensive harm. This difficulty bears out in some of the existing analyses of the Norman case. For example, Ferzan claims that the kidnapping paradigm is sufficient in Norman's case to justify defensive killing. By imprisoning Norman in her home, Ferzan claims, Norman's abuser is "continuingly invading the victim's rights." But Ferzan's argument for this claim is ambiguous.

> Judy Norman married J. T. Norman when she was . . . but fourteen years of age. She could never escape him. During her time in captivity, she was "forced into prostitution"—that is, Judy Norman was repeatedly raped. . . . There was not an imminent or inchoate threat of death. There was an ongoing and continuing denial of life. . . . If it is in fact the case that the battered woman can show that she is a hostage and cannot escape her husband, this alone should be sufficient for the exercise of deadly force.[20]

There seem to be two distinct claims in this quotation. The first is that the repeated, violent, and life-denying attacks that Judy Norman's abuser imposed upon her and would continue to impose upon her were sufficient to justify killing in self-defense. The second is that the imprisonment alone—the fact "that she is a hostage and cannot escape"—is sufficient to justify killing in self-defense. Ferzan is here claiming that the many instances of threats of imminent, though nonlethal, discrete harm are sufficient to justify Norman's defensively killing her abuser. At the same time, Ferzan also claims that the threat of continuing harm—the fact that "she is a hostage and cannot escape"—is sufficient to justify Norman's defensively killing her abuser. These intuitions about sufficient

[20] This quotation and those in the next paragraph are taken from Ferzan, "Defending Imminence," 253–55.

conditions for lethal defensive harm are similar to the intuitive responses to Trolleyology and Home Invaders. We have a strong intuition that defenders may cause harm to defeat the aggression, but it is not clear which kind of threat is doing the justificatory work. Looking only at Norman's case will not provide an answer. But if threats of continuing harm can justify defensive killing—as in Pleasant Detention—then they can also justify defensive killing in cases that include multiple kinds of threat—as in Domestic Abuse. Depending on the circumstances, and especially on the magnitude and duration of the imprisonment, the continuing threat can be independently sufficient to justify lethal defensive harm.

3. Continuing Harms in War

The relationship between threats of continuing harm and proportionality is relevant, not just to cases involving individuals, but to cases of political conflict as well. Reductivist (sometimes called "revisionist") just war theory considers justified killing in war to be an instance of justified defensive harming. This might be easy to conceptualize in a firefight, for example, in which unjust combatants wrongfully threaten to kill just combatants.[21] Just combatants have self-defense justifications for shooting and other-defense obligations to shoot. But the experience of war is often characterized by the perennial waiting that takes place between the firefights—the "months of boredom punctuated by moments of terror."[22] Can just combatants be justified in seeking out and engaging unjust combatants even while the unjust combatants are merely in the midst of this waiting? It would be counterintuitive, perhaps even paradoxical, to hold that just combatants are justified in going to war and justified in returning fire during the firefight and yet not justified in seeking out enemy combatants to defeat them. Most just war theorists hold that combatants on the just side are morally justified in seeking out enemy combatants to try to kill them. I argue here that the account of threats of continuing harm I have developed above can, under some circumstances, help to explain that moral justification.

To be sure, the conception of war as an ongoing activity, or of a soldier's discrete actions as constitutive of an ongoing conflict, is common in the just war literature. For example, Walzer considers the case of a soldier behind enemy lines. Walzer suggests that perhaps the soldier in his tent is "smoking his morning cigarette, [and] thinking only of the coming battle and of how many of

[21] Throughout, I refer to combatants on the just side of a conflict as "just combatants," and combatants on the unjust side as "unjust combatants."

[22] A Young British Officer, "The Baptism of Fire," 979.

his enemies he will kill. He is engaged in war-making just as I am engaged in writing this book; he thinks about it all the time or at the oddest moments."[23] Walzer goes on to reject this notion on the grounds that most combatants do not approach war in this way. Even if this description of the soldier in his tent is the correct one, though, the fact that the soldier is wholly devoted to his cause probably bears on his culpability; but it is not clear that by thinking about his next battle he is somehow engaged in a threat of harm. McMahan uses similar language in his response to Rodin's conception of the bloodless invasion. "War involves threats that consist of activities organized in phases over extended periods of time. A soldier sleeping in invaded territory has already attacked and is engaged in attacking in the same way that I am engaged in writing this essay even while I pause to make a cup of tea."[24] This is a better analogy. It is not that a soldier at rest is thinking a lot about the next battle, but that he is engaged, not just in discrete battles, unrelated to one another, but in a war. And on these grounds, we can be confident that he will cause harm again in the future.

These descriptions, though they incorporate a sense in which threats are ongoing—a soldier constantly threatens to cause harm at some discrete moment in the future—they refer only to the lethal actions a soldier has taken or will take in the future and not to threats of continuing rights violations. According to the distinction in the introduction above, they are continuing threats of discrete harm, but they are not threats of continuing harm. Limiting one's conception of threats in war only to imminent threats of discrete harm and continuing threats of discrete harm is like evaluating Judy Norman's case with reference only to the constant threat of discrete harm her abuser posed without considering the fact that she was imprisoned in her home. The picture of the morally relevant threats and harms is incomplete. War often involves threats of continuing rights violations because, as Brian Orend has put it, war is about "governance itself." Armies do not meet on the field of battle as street gangs, each fighting for its own narrow interests. Rather, armies meet as representatives of political communities and use violence to "resolve disputes over governance."[25] If wars are fought over governance, then at least some wars will entail the violation of one group's right to self-governance. As I argue below, invasion and occupation can, under some circumstances, be considered continuing harms alongside imprisonment and kidnapping.

[23] Walzer, *Just and Unjust Wars*, 143.

[24] McMahan, "War as Self-Defense," 76. It strikes me that both Walzer and McMahan compare academic writing to war. Though the comparison might be strained, one sympathizes with the analogy.

[25] Orend, *The Morality of War*, 4.

In this section, I introduce three combat cases to show how the conception of continuing harm applies to just war theory.

3.1. Sleeping Soldier

The sleeping soldier might pose a threat, but what kind of threat does he pose? Consider this adaptation of Robert Graves's account from the First World War.

> *Sleeping Soldier*: I saw a German . . . through my telescopic sights. He was [sleeping] in the German third line. . . . I handed the rifle to the sergeant with me. "Here, take this. You're a better shot than I am." He got him; but I had not stayed to watch.[26]

I take it for granted that Graves and his sergeant did have a moral justification, but on what grounds? If Graves's sergeant has a self- or other-defense justification for killing the sleeping soldier, against what threat does he defend? The sleeping soldier certainly does not pose an imminent threat of harm while he sleeps. He probably does pose a nonimminent threat of discrete harm in that, eventually, he will pick up his weapon and fire upon the adversary. But there is more to the sleeping soldier's liability to harm than the nonimminent threat of discrete harms he poses.

On the account of threats of continuing harm developed above, it might matter morally where the sleeping soldier is sleeping. For instance, Graves and his sergeant happened upon this particular German soldier in German-occupied Cuinchy, France. Even as the sleeping German soldier posed a nonimminent threat of lethal harm, he also caused a continuing harm. There, in the German third line in Cuinchy, he continually violated, and at once continually threatened to violate further, the French people's right to political autonomy and to territorial integrity. Perhaps this second kind of harm and second kind of threat are not necessary for Graves's sergeant's moral justification for killing the German soldier. In other words, perhaps, even if the German soldier threatened continuing harm, the German soldier's nonimminent threat of discrete lethal harm is sufficient to justify killing him. Even so, the proportionality calculus is at least theoretically sensitive both to threats of discrete harms as well as to threats of continuing harm. As we shall see below, in some cases, distinguishing between these various kinds of threat can have an effect on liability and the moral justification for killing in war.

[26] I have modified this case from a bathing soldier to a sleeping soldier to sidestep some of the psychological questions that arise in the bathing case. See Graves, *Goodbye to All That*, 112; Walzer, *Just and Unjust Wars*, 140.

3.2. Bloodless Invasion

The just war literature is well acquainted with the idea that soldiers on the unjust side can threaten lesser interests without directly threatening vital interests. Often called "the bloodless invasion," the idea that one state can begin a takeover of a neighboring state's sovereign territory is not merely theoretical.

In March 2014, BBC correspondent John Simpson attempted to pass from the mainland of Ukraine into the Crimean Peninsula. In the preceding weeks, the Russian government had steadily increased the Russian military presence on its bases in Crimea—bases it was permitted to use under bilateral agreement with Ukraine. By the first of March, Russian forces had established a new checkpoint at Armyansk, and armed men in fabricated Ukrainian army and police uniforms detained Simpson and his colleague at the checkpoint while they searched the reporters' equipment and confiscated much of it. When the two men were allowed to pass through the checkpoint, a man in a Ukrainian police uniform called out, "Welcome to Russia!" Simpson writes,

> The Annexation of Crimea was the smoothest invasion of modern times. It was over before the outside world realized it had even started. And until Tuesday 18 March [when one Ukrainian was killed and another injured], it was entirely bloodless.[27]

The idea of a bloodless invasion was a useful thought experiment long before the Russian invasion of Crimea. It is a useful philosophical case because it separates individual citizens' rights to life and the state's putative right to territorial integrity. Most unjust invasions in the real world involve threats both to territorial integrity and to people's lives. In these cases, as in Sleeping Soldier case above, reductivist accounts of just war theory can justify the killing of invading combatants on individualist defensive harming grounds. The aggressors threaten the defenders' rights not to be killed, and defenders justifiably and defensively kill aggressors, independent of any appeal to putative rights of the political community writ large.

Here is a generic version of the case.

[27] Simpson, "Russia's Crimea Plan Detailed, Secret and Successful." Lazar admits that the bloodless invasion thought experiment might be strictly hypothetical without any real-world instance. But he also presciently suggests that "states like Russia and China, which have long-simmering territorial disputes with their neighbors, [might] take advantage of the opportunity to settle those disputes through bloodless invasion." The chapter in which he wrote these words was published just one month before the 2014 Russian invasion of Crimea. Lazar, "The Problem of Political Aggression," 24.

> *Bloodless Invasion:* Soldiers of the state of Northland unjustly invade the democratic republic of Southland. The Northlandic soldiers are heavily armed and tell the people of Southland that they intend only to replace the Southlandian government with their own officials and autocratic institutions. They do not intend physically to harm any Southlandian citizens. But if Southlandian citizens or soldiers resist, the Northlandic soldiers will use lethal force to subdue them.

At first, Bloodless Invasion might seem like a fabricated, philosophical case. But there are reasons—even apart from the 2014 Russian action in Crimea—to think it is plausible. The fact that the threat of harm to vital interests is conditional is one such reason. The invading force tells would-be defenders that they will be harmed only if they resist. It might be the case that a great many unjust invasions throughout history would have been bloodless had they not been resisted. It is at least theoretically possible that many unjust invasions would be bloodless were it not for the resistance of the victims.[28]

The soldiers of Northland violate Southland's territorial integrity and political sovereignty. According to international law and to collectivist accounts of just war theory, Northland's violation of Southland's state rights amounts to a just cause to wage a defensive war against Northland. This war will undoubtedly include not only lethal harms to Northland's soldiers, but also the unintentional deaths of numerous noncombatants, perhaps on both sides of the conflict. In Bloodless Invasion, at first glance, it looks as though the reductivist must conclude that killing the bloodless invaders violates the principle of proportionality. If Southlandian soldiers kill Northlandic soldiers to defend their rights of, say, political autonomy, that lethal harm will *ex hypothesi* violate the proportionality principle in the narrow sense. In Rodin's words, "The argument from bloodless invasion is designed to show that defending the lives of citizens is not a necessary condition for national-defense."[29] The bloodless invasion thought experiment pressures reductivists either to admit that states have no rights of political autonomy and territorial integrity or that, if states do have those rights, they are not justified in causing lethal harm to defend them. Just as Graves's sergeant had to decide whether he was justified in killing a soldier who did not pose a threat of imminent, lethal harm, Southlandians must decide if they are justified in waging a war to defeat an enemy that does not pose a threat of imminent, lethal harm.

[28] Both Norman and Lazar have made this point. Norman, *Ethics, Killing, and War*, 135; Lazar, "The Problem of Political Aggression," 19.

[29] I follow McMahan's use of "narrow" and "wide" proportionality. McMahan, *Killing in War*, 20–21. For the Rodin quotation, see *War and Self-Defense*, 131.

Do the Northlandic soldiers threaten the vital interests of Southlandians, or do they threaten only lesser interests? Recall that vital interests are those interests that one is permitted to defend with lethal harm. Lesser interests may also be defended, and indeed one may still cause harm to defend them, but one may not cause lethal harm in their defense.[30] By invading Southland and taking over its political institutions, Northlandic soldiers do in fact violate the Southlandian citizens' rights to political autonomy. In addition, they pose a conditional threat—one instance of a threat of nonimminent harm—against the Southlandian citizens' lives. The first question to ask is about necessity. Is killing Northlandic soldiers the least morally weighted moral harm Southlandian soldiers can cause to defeat the unjust threat? If so, then the necessity condition has been met. The second question is about proportionality. Is it permissible for Southlandian soldiers to kill Northlandic soldiers, not in defense of their rights to life, but in defense of their rights to political autonomy? In other words, may soldiers on the defensive side kill to defend a *prima facie* lesser interest? Without a conception of threats of continuing harm, Southlandian soldiers lack a moral justification to cause lethal defensive harm because the proportionality condition has not been met. And if so, then we must conclude that, whatever rights states have to territorial integrity, they are not morally justified in waging war to defend those rights. Though this may turn out to be the correct conclusion, it is antithetical to common sense intuitions about the rights of states as well as long-standing international norms.[31]

There are two ways in which previous accounts of reductivist just war theory have responded. My argument is that the conception of threats of continuous harm provides a more complete response to the bloodless invasion challenge than do either of these two options. The first appeals to escalation and the second to interpersonal aggregation. Space does not permit a thorough analysis here, but a sketch of each approach will be helpful. McMahan has developed an appeal to escalation according to which, when the Northlandic soldiers unjustly and bloodlessly invade Southland, they threaten Southlandian citizens' lesser interests but also pose a conditional threat of nonimminent harm against their vital interests if they resist. If Southlandian soldiers do resist, Northlandic soldiers will pose an unjust threat to vital interests and, at that point, Southlandian soldiers will be justified in causing lethal defensive harm. Thus, *ex hypothesi*, Southlandian soldiers are justified in causing lethal harm to defend against the bloodless invasion. McMahan ultimately rejects this account on grounds

[30] See, respectively, Rodin, "The Myth of National Self-Defence," 80–81; McMahan, "War as Self-Defense," 77–79, and "What Rights May Be Defended by Means of War?," 126.

[31] For example, Article 2(4) and Article 51 of the UN, "Charter of the United Nations."

that, even if Southlandian soldiers are permitted to accept additional risk to themselves by escalating, they are not permitted to accept the additional risk to their fellow citizens that would undoubtedly result from the escalation.[32]

The second reductivist response to Bloodless Invasion appeals to interpersonal aggregation. On this line of thought, even if a Southlandian soldier would be unjustified in killing to defend her own lesser interest, she acts not merely for herself but on behalf of her fellow citizens. Once aggregated across members of the political community, the sum of the Northlandic soldiers' violations of lesser interests against which Southlandian soldiers defend on this view meets the proportionality requirement in both its narrow and wide senses.[33] The challenge facing this argument is that war will most likely result in harms to citizens of the aggressor state, too. These harms must also be aggregated in the proportionality calculus. In other words, if the number of Southlandians who will be harmed weighs on one side of the proportionality calculus, surely the number of Northlandians who will be harmed must weigh on the other side.[34]

What is missing from both responses to the bloodless invasion challenge is any sense of duration. Bloodless Invasion constitutes a continued denial of persons' rights to political autonomy over a long period of time. Whether Northlandic soldiers pose a continuing threat to vital or lesser interests depends both on the magnitude and the duration of the threatened harm. Will Northland occupy and dictate to Southlandians for a month? For a year? Forever? As was the case in Pleasant Detention, if the Northlandic soldiers continually violate the rights of Southlandians, the proportionality calculus is sensitive to duration. There is some threshold in time at which the Northlandic violation of Southlandian interests crosses over from being a violation of lesser interests to becoming a violation of vital interests. In other words, in the war between Northland and Southland—a war over governance—on the most fundamental level, Northland poses a threat of continuing harm to Southlandians by violating their rights of political autonomy. In so far as wars are violent means of resolving disputes over governance, this is true in the general case: in war generally, aggressors continually violate and continually threaten to violate victims' rights of self-governance—even when they also threaten to violate rights of life. Even in a bloodless invasion, therefore, Southlandians are morally justified in

[32] McMahan defended this view in his earlier work. In more recent work, he provides a helpful and clear account of the argument from escalation, but ultimately rejects it. McMahan, "Innocence, Self-Defense and Killing in War," 195–96, and "What Rights May Be Defended by Means of War?," 147–48.

[33] For an example of this kind of argument, see Frowe, *Defensive Killing*, 139–43.

[34] For a critique of this kind, see Lazar, "The Problem of Political Aggression," 32. For Frowe's response, see *Defensive Killing*, 139.

using lethal force to defend against the threat of continuing harm as long as the harm would otherwise continue for a sufficient duration.

The proportionality calculus in Bloodless Invasion depends on the continuing harm Northland threatens, aggregated over time—this is the intertemporal aggregation of harm. On the surface, it looks as though intertemporal aggregation of harm and intrapersonal aggregation of harm are synonymous. Intrapersonal aggregation is not aggregation across persons, but within a single person's life. But normally, the intrapersonal aggregation of harm refers to the additive harm a single person suffers if she is the victim of multiple discrete harms over time. As Frowe puts it, additive harm is aggregated interpersonally if I break lots of different people's arms. The additive harm is aggregated intrapersonally if I break the same person's arm lots of times. But the kinds of continuing harms I have in mind—kidnapping, imprisonment, enslavement—are dissimilar from Frowe's case of a repetitive series of discrete harms to the same person. In between discrete acts of arm-breaking, I do not harm the victim. But cases of kidnapping, enslavement, and imprisonment, as well as cases of invasion and occupation, include both the threat of denying and the actual denial of victims' freedoms.[35]

It is possible to preserve the conceptual distinction I apply here if we consider intertemporal aggregation to be a species of the broader genus of intrapersonal aggregation. This is etymologically appropriate, given that the continuing harm measured over time (intertemporal) is harm caused to the same person (intrapersonal). The point I make here is only that causing continuing unjust harm and causing repetitive unjust harm should remain conceptually distinct.

3.3. *Anwar al-Aulaqi*

In the two cases above, combatants pose a nonimminent threat of lethal harm and cause continuing harm. But how should we interpret a case in which a combatant poses a nonimminent threat of lethal harm without causing continuing harm? In other words, what might the moral difference be if a combatant is engaged in hostilities, but is not engaged in invasion, occupation, or otherwise threatening the political autonomy of his victims? The targeted killing of Anwar al-Aulaqi might prove to be such a case.

The US targeting of Anwar al-Aulaqi has received significant attention, most notably because Aulaqi was a US citizen—a fact that raises important questions about citizenship in political philosophy and in US law. But there are other facts about the Aulaqi case that pertain to threats of continuing harm. Assume for

[35] For Frowe's description, see *Defensive Killing*, 140. For a discussion of the role of interpersonal aggregation in proportionality, see Tadros, "Past Killings and Proportionality in War."

the sake of argument that Aulaqi was an operational leader within al-Qaeda.[36] Although Aulaqi did pose a threat to US citizens, he did not invade the United States, having departed in 2007, never to return. Al-Qaeda's war against the US was carried out in discrete terrorist attacks and not on violations of political autonomy. In other words, Aulaqi, and other al-Qaeda leaders like him, posed nonimminent threats to US persons—but he did not cause continuing harm.

The Obama administration's official position was that Aulaqi was a legitimate military target because, in his role as an operational leader within al-Qaeda, he posed a "continuing and imminent threat" of violent attack against US persons and interests, but this phrasing is misleading. Suppose that from the use of "violent attack" we can infer that Aulaqi threatened American citizens' vital interests. What can it mean for a threat to be both imminent and continuing? Though it is not with reference to the Aulaqi case, Ferzan insists that "always imminent" is a contradiction in terms. If so, the same contradiction must obtain in the supposed "imminent and continuing" threat Aulaqi posed. If an imminent threat is one that will cause harm "in an instant" or "at once," how is it possible that Aulaqi posed an imminent threat for a year and a half?[37] Once threats of continuing harm are properly understood, it is plausible that an unjust combatant can simultaneously pose some combination of threats of imminent harm, threats of traditional nonimminent harm, and threats of continuing harm. But, in Aulaqi's case, even though he posed threats of discrete harm, he did not pose a threat of continuing harm.

This case helps to show that the application of threats of continuing harm to the morality of war is sensitive to whether the war, campaign, or battle in question amounts to an invasion or an occupation. Like the sleeping soldier, at the time Aulaqi was killed, he was not actively engaged in a firefight. However,

[36] The public was originally made aware of Aulaqi's online presence, pro al-Qaeda views, and propagandist role after reports that Aulaqi had exchanged emails with Nadil Malik Hasan prior to Hasan's 2009 attack on Fort Hood, a US Army base in Texas. Shortly thereafter, however, the US government claimed that Aulaqi was no mere propagandist and was in fact a member of al-Qaeda and an operational leader within that organization. Specifically, the US government claimed that Aulaqi recruited, trained, and directed Umar Farouk Abdulmutallab, the so-called Underwear Bomber whose attempt to bring down a Detroit-bound airliner was thwarted on Christmas Day 2009. Epstein, "The Curious Case of Anwar Al-Aulaqi," 725; Chesney, "Who May Be Killed?," 9.

[37] For the US government's use of "continuing and imminent," see Office of Legal Counsel, "Memorandum for the Attorney General Re: Applicability of Federal Criminal Laws and the Constitution to Contemplated Lethal Operations against Shaykh Anwar Al-Aulaqi;" Holder, "Attorney General Eric Holder Speaks at Northwestern University School of Law." For the Ferzan quotation, see "Defending Imminence," 229. For the factual claims about the Aulaqi case, see Epstein, "The Curious Case of Anwar Al-Aulaqi," 2; Van Schaack, "The Killing of Osama Bin Laden and Anwar Al-Aulaqi," 6; Shane, *Objective Troy*.

unlike Sleeping Soldier and Bloodless Invasion, Aulaqi has not unjustly invaded the territory of a political community and does not seem to pose a threat of continuing harm. Most unjust wars include threats to life and to political autonomy. Bloodless Invasion is a helpful case because it includes the latter but not the former. Aulaqi's case is different in that it includes the former but not the latter. If it is the case that threats of continuing harm—especially those harms that will continue for a long time—help to justify harms caused in defensive wars, there might be an important difference between the defensive harms that are justified to defeat invasions or occupations and the defensive harms that are justified to defeat threats that are not invasions or occupations. The conception of threats of continuing harm for which I have argued might cast doubt on *jus ad bellum* justifications in any cases in which the enemy does not credibly threaten political autonomy—and more specifically, in twenty-first-century US and coalition asymmetric wars against transnational terror organizations. But a thorough investigation into this question falls outside the scope of this paper.

4. Conclusion

At different times, attackers can pose imminent threats, nonimminent threats, continuing threats, and combinations thereof. For example, when soldiers on patrol discover opposing soldiers, they will pose a threat of imminent harm to those soldiers. Threats of conditional, nonimminent harm are more common. In fact, most combatants pose a conditional threat of nonimminent harm most of the time. Almost any combatant in the theater of combat operations—even noncombat troops such as maintenance, supply, or communications soldiers—will return fire if fired upon. But unjust combatants also pose threats of continuing harm when they are in the act of violating either civilians' or just combatants' rights. Whether defenders are justified in causing lethal defensive harm depends upon, *inter alia*, the magnitude and the duration of the continuing harm imposed.

The just combatant's justification for killing the sleeping soldier is not necessarily grounded in a threat of imminent harm. By participating in an unjust invasion, the sleeping soldier is in the act of violating, or attempting to violate, rights of liberty and political autonomy. The significance of these violations may seem trivial, but there are open questions about their duration. If the sleeping soldier's state had intended to affect permanent change—that is, to violate rights of political autonomy indefinitely—then the sleeping soldier is relevantly similar to Agatha in Pleasant Detention. The same is true of Bloodless Invasion. If the Russian soldiers in 2014 intended not only to invade Crimea, but also to keep it indefinitely—a conclusion that has become undeniable following

Russia's 2022 invasion of Ukraine—it is plausible that they violate the rights of Ukrainian citizens and that they intend to do so indefinitely.

There might be some outlying cases in which violent resistance to an unjust invasion violates narrow proportionality. One thinks, for example, of the Soviet Union's annexation of Finland during the Second World War. Imagine—even if doing so gives Stalin more credit than is due—that the Soviet Union had every intention to use Finnish territories unjustly as a buffer to deter German aggression only temporarily and to return the territories to Finland after a few years. The Soviet Union might have intended only a bloodless invasion. "Do you want matters to lead to a conflict?" the Soviet foreign minister asked Finnish negotiators. In the face of Finnish implacability, the same minister closed the negotiations by saying, "We civilians can see no further in the matter; now it is the turn of the military to have their say."[38] The annexation of Finnish territories, therefore, has all the hallmarks of a bloodless invasion, including the conditional threat of lethal harm. The Finns responded with defensive force, and the Winter War had begun. Surely the Soviet annexation of Finnish territory was unjustified and Soviet soldiers posed a threat of continuing harm to the Finnish people's rights to freedom, political autonomy, and property even before they posed threats of lethal harm. But suppose the duration of those violations was temporally limited. Suppose it would have been only for a couple of years, or even a couple of months. It is at least theoretically possible that the Soviet annexation of Finland—had it not been resisted with violence—would have failed to reach the magnitude and duration thresholds to justify lethal defensive force.

Even if it is the case that Finland's violent resistance to Soviet invasion in 1939 violated the proportionality principle in the narrow sense—and this claim is dubious—cases like this one are exceedingly rare. Invading forces often violate persons' rights to political autonomy and to property and threaten to do so for a very long period of time. If war is a conflict over governance, then demands for changes in governance are not likely to be demands only for a temporary change. And if the invasion—bloodless though it may be—threatens a permanent violation of otherwise lesser interests, then just as Violet is justified in killing Agatha in Pleasant Detention, so are defenders justified in warding off the bloodless invasion with lethal harm.[39]

United States Air Force
joseph.chapa@us.af.mil

[38] Sechser, "Goliath's Curse," 645–47.

[39] I am grateful to Cécile Fabre, Tom Simpson, Jeff McMahan, and B.J. Strawser for the instructive comments they provided on early versions of this paper. I am also grateful to the journal's anonymous reviewers whose comments likewise strengthened the paper.

REFERENCES

Belew, Christine M. "Killing One's Abuser: Premeditation, Pathology, or Provocation?" *Emory Law Journal* 59, no. 3 (2010): 769–808.

Chesney, Robert. "Who May Be Killed? Anwar Al-Awlaki as a Case Study in the International Legal Regulation of Lethal Force." In *Yearbook of International Humanitarian Law*, vol. 13, 3–60. Berlin: Springer, 2010.

Epstein, Michael Robert. "The Curious Case of Anwar Al-Aulaqi: Is Targeting a Terrorist for Execution by Drone Strike a Due Process Violation When the Terrorist Is a United States Citizen?" *Michigan State University College of Law Journal of International Law* 19 (2011): 723–44.

Fabre, Cécile. *Cosmopolitan War.* Oxford: Oxford University Press, 2012.

———. "War Exit." *Ethics* 125, no. 3 (April 2015): 631–52.

Fabre, Cécile, and Seth Lazar, eds. *Morality of Defensive War*. Oxford: Oxford University Press, 2014.

Ferzan, Kimberly Kessler. "Defending Imminence: From Battered Women to Iraq." *Arizona Law Review* 46 (2004): 213–62.

Fletcher, George P. "Domination in the Theory of Justification and Excuse." *University of Pittsburgh Law Review* 57 (1996): 553–78.

Frowe, Helen. *Defensive Killing.* Oxford: Oxford University Press, 2014.

Graves, Robert. *Goodbye to All That.* London: Penguin, 2000.

Holder, Eric. "Attorney General Eric Holder Speaks at Northwestern University School of Law." Office of Public Affairs, US Department of Justice, March 5, 2012. https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-northwestern-university-school-law.

Lazar, Seth. "National Defence, Self-Defence, and the Problem of Political Aggression." In Fabre and Lazar, *The Morality of Defensive War*, 11–39.

Lewis, C. S. *The Lion, the Witch and the Wardrobe.* New York: Collier Books, 1970.

McMahan, Jeff. "Innocence, Self-Defense and Killing in War." *Journal of Political Philosophy* 2, no. 3 (September 1994): 193–221.

———. *Killing in War.* Oxford: Oxford University Press, 2009.

———. "Proportionality and Time." *Ethics* 125, no. 3 (April 2015): 696–719.

———. "War as Self-Defense." *Ethics and International Affairs* 18, no. 1 (March 2004): 75–80.

———. "What Rights May Be Defended by Means of War?" In Fabre and Lazar, *The Morality of Defensive War*, 115–56.

Norman, Richard. *Ethics, Killing, and War.* Cambridge: Cambridge University Press, 1995.

Office of Legal Counsel. "Memorandum for the Attorney General Re:

Applicability of Federal Criminal Laws and the Constitution to Contemplated Lethal Operations against Shaykh Anwar Al-Aulaqi." US Department of Justice. July 16, 2010. https://www.justice.gov/sites/default/files/olc/pages/attachments/2015/04/02/2010-07-16_-_olc_aaga_barron_-_al-aulaqi.pdf.

Orend, Brian. *The Morality of War.* 2nd ed. Peterborough, ON.: Broadview Press, 2013.

Rodin, David. "The Myth of National Self-Defence." In Fabre and Lazar, *The Morality of Defensive War*, 69–89.

———. *War and Self-Defence.* Oxford: Oxford University Press, 2002.

———. "The War Trap: Dilemmas of *Jus Terminatio*." *Ethics* 125, no. 3 (April 2015): 674–95.

Rosen, Richard A. "On Self-Defense, Imminence, and Women Who Kill Their Batterers." *North Carolina Law Review* 71, no. 2 (1993): 371–411.

Sechser, Todd S. "Goliath's Curse: Coercive Threats and Asymmetric Power." *International Organization* 64, no. 4 (October 2010): 627–60.

Shane, Scott. *Objective Troy: A Terrorist, a President, and the Rise of the Drone.* New York: Tim Duggan Books, 2016.

Simpson, John. "Russia's Crimea Plan Detailed, Secret and Successful." BBC News, March 19, 2014. https://www.bbc.com/news/world-europe-26644082.

Singer, Peter. "Bystanders to Poverty." In *Ethics and Humanity: Themes from the Philosophy of Jonathan Glover*, edited by N. Ann Davis, Richard Kershen, and Jeff McMahan, 185–201. Oxford: Oxford University Press, 2010.

Tadros, Victor. "Past Killings and Proportionality in War." *Philosophy and Public Affairs* 46, no. 1 (Winter 2018): 9–35.

United Nations. "Charter of the United Nations." *Yale Law Journal* 55, no. 5 (August 1946): 1291–317.

Van Schaack, Beth. "The Killing of Osama Bin Laden and Anwar Al-Aulaqi: Uncharted Legal Territory." *Yearbook of International Humananitarian Law* 14 (2011): 255–325.

Walzer, Michael. *Just and Unjust Wars: A Moral Argument with Historical Illustrations.* 5th ed. New York: Basic Books, 2015.

A Young British Officer. "The Baptism of Fire." In *New York Times Current History of the European War*. New York: The Times, 1914.

Journal of Ethics and Social Philosophy
vol. 26, no. 1 · november 2023

https://doi.org/10.26556/jesp.v26i1.2602

AGAINST DEFERENCE TO AUTHORITY

Travis Quigley

Joseph Raz's service conception of authority retains significant influence in moral, political, and legal theory. I raise a problem for the theory and suggest a significant revision in response. Many commentators and critics have focused on whether the service conception fits with the concept of authority or law.[1] The objection I raise lies, instead, in the justificatory structure of Raz's view, which perhaps explains why it has gone largely unseen for so long. In short, I argue that there is a deep tension between three core components of the service conception: that authority is justified piecemeal to each subject depending on their epistemic situation; that within its piecemeal domain, authority provides exclusionary reasons to obey; and that authority features directly in practical reason.

Each of these claims represents a core part of the appeal of the theory. The piecemeal nature of authority is one of Raz's principal innovations, allowing the service conception to sidestep the arguments of philosophical anarchists that no state can create general (even if defeasible) reasons to obey. The exclusionary power of authority reflects a commonly held conceptual feature of authority, that it is decisive, somehow akin to the parent commanding the child or the military officer commanding the private. The role of authority in practical reason enables authority to provide a service; if authority were merely an abstract feature that obtains or does not, it would not be able to help subjects comply with reason.

Laws are necessarily coarse-grained, operating at a level of generality that allows practical functionality. This means that even the best states will routinely make particular suboptimal commands. Raz allows for state errors and makes some room for them in practical reason by excepting from authority any epistemic domains in which a subject is an expert and thus need not rely on the state to comply with reason. But this is not sufficiently piecemeal, as I will show. It is possible to identify state errors even when one is not an expert,

[1] For just a handful of papers from the large literature, see Darwall, "Authority and Reasons"; Enoch, "Authority and Reason-Giving"; Hershovitz, "The Role of Authority"; Perry, "Second-Order Reasons"; and Raz, "Revisiting the Service Conception."

just by having some special information pertinent to a particular application of a law. Call this *accidental expertise*. This means there is no *ex ante* specifiable domain for piecemeal authority. Epistemic situations vary case by case, as well as agent by agent.

This opens up an inconsistency. If authority is exclusionary, the lack of specifiable domains of expertise leaves subjects in the lurch. They have to decide whether they are under the authority of the state in a particular situation where this decision requires deliberating about their degree of confidence in whatever information they happen to have. But this kind of practical deliberation about whether a law is worth obeying in a particular circumstance is precisely what exclusionary reasons are meant to rule out. If subjects cannot know *ex ante* whether authority obtains, and they cannot deliberate effectively without presupposing that authority *does not* obtain, then authority cannot function in practical reason because its precise scope cannot be known.

Raz has recently discussed the "knowability condition" on authority more explicitly.[2] As he explains there, "The point of being under an authority is that it opens a way of improving one's conformity with reason." This is central to Raz's entire account of authority.[3] Authority cannot improve conformity with reason if the scope of authority cannot be known. So I do not take seriously the possibility of eliminating the knowability condition. Eliminating the piecemeal nature of authority is an obvious nonstarter. Instead, I propose that the service conception should drop exclusionary reasons, and I provide an alternative.[4]

I will call the alternative *habitual obedience*. The relevant notion of habit is a *trainable but automatic* disposition to act on an established pattern or routine. Habits lie on a spectrum of dispositions to act: to one side lie instincts, which are not (significantly, in normal circumstances) trainable; to the other side lie principles, which are not (significantly) automatic. Automatic dispositions risk, and indeed accept, certain inevitable mistakes. To rely on an automatic process necessarily means being blind to some countervailing reasons that might be noticed upon reflection. The corresponding benefit is *fluency*—automatic habits save time, allow fluid and natural responses to circumstances, and can mitigate the influence of biases.[5]

I argue that a habit of obedience to (legitimate) law is a superior disposition, by Raz's own justificatory lights, compared to treating the law as creating

[2] Raz, "Revisiting the Service Conception," 1025.

[3] Most obviously in the core Normal Justification Thesis, discussed below.

[4] I do not claim that this is the only possible alternative.

[5] See Railton, "Practical Competence," for discussion of fluency, as well as Pettit, "The Inescapability of Consequentialism"; Pollard, "Can Virtuous Actions Be Both Habitual and Rational?"

exclusionary reasons to obey. This is compatible with the other core features of the service conception. Justified habits of obedience will vary widely between individuals, and the law on this view still aids citizens by helping them better conform to reason than they could by deliberating on their own. But it is not clear that this is any longer a service conception of *authority*, rather than (merely?) a service conception of *law*. Raz takes exclusionary reasons to be constitutive of authority, and it is intuitive that authority connotes decisiveness within some domain. But habits do not have any particular specifiable domain: they are trained instead in the *normal case* and are delimited by "intervention control" (Pollard) or "red lights" (Pettit), which cue us to cut off the automatic process and undertake conscious deliberation.[6] The scope of habits, while better or worse justified depending on how well they serve us, is not itself rationally cognizable. Whether one views the use of habits as compatible with a revisionary stance on authority or instead as a form of skepticism about authority is, to some extent, a matter of taste. But that is where the other central—and very appealing—elements of the Razian approach lead.

Here is the plan. Section 1 discusses the role of coordination in Raz's conception of the practical authority of law. This is a preliminary argument explaining that while coordination makes political authority distinctively practical, deferring to political authority (by treating it as exclusionary) is only justified if it *also* is the best strategy for complying with reasons in a manner that is highly similar to theoretical authority.[7] I suggest that the coordinative role of authority is exactly what leads to the problems with exclusionary deference in the political case. Section 2 develops the costs of deference to the law. The basic strategy is to develop several examples and then argue that Razian strategies to avoid the examples run afoul of the knowability condition. Section 3 develops the habitual obedience strategy and its advantages over exclusionary deference.

1. Coordination and Practical Authority

To accept an authority as binding is to treat it as creating exclusionary reasons to obey. On Raz's view, an authoritative command generates both a first-order reason to obey and a second-order reason not to act on—to exclude—at least some possible reasons for disobedience.[8] This conjunction is called a

[6] Pollard, "Can Virtuous Actins Be Both Habitual and Rational?"; Pettit, "The Inescapability of Consequentialism."

[7] As Raz himself emphasizes ("Revisiting the Service Conception," 1033–34).

[8] It does not preclude *thinking* about reasons for disobedience, so long as those reasons are not acted upon. See section 3.2 below. It does seem to imply at least a *permission* not to consider countervailing reasons since such reflection is practically idle.

preemptive reason.[9] Authority is justified on the basis of the Normal Justification Thesis:

> *Normal Justification Thesis* (NJT): The normal way to establish that a person has authority over another person involves showing that the alleged subject is likely better to comply with reasons that apply to him (other than the alleged authoritative directives) if he accepts the directives of the alleged authority as authoritatively binding and tries to follow them, rather than by trying to follow the reasons that apply to him directly.[10]

There are two main rationales for how the law can satisfy the NJT. One is epistemic: the law is formulated by experts, and individuals are prone to error. This rationale is similar to deference to theoretical authorities, which create preemptive reasons on the same basis.[11] The second is coordinative: the exclusionary character of the law allows everyone to safely act on the assumption that all other subjects will comply. The main discussion of the epistemic rationale comes in section 2. The point of this section is to head off the possibility that the special practical nature of political authority insulates it from the arguments I give there. On the contrary, the practical nature of authority is precisely what opens it up to my objection.

Here are two quick arguments for the conclusion that, while epistemic and coordinative considerations may be "inextricably mixed" for political authorities, this mixture must *contain* a robust epistemic endorsement of deferring to the state on the basis of its expertise.[12]

The first argument is that if this were not so, it would appear that coordination on its own is sufficient for authority. This would seem to take up a Hobbesian rather than a Razian line.[13] Achieving coordination, despite its great value, is clearly insufficient for exclusionary authority. A tyrannical regime that rules by the iron fist of harsh punishment can achieve coordination, at least some of which will be beneficial compared to the state of nature. But such a regime is not authoritative. Pernicious regimes can also establish coordination without force through the sufficient development of state ideology. If the citizens of a state freely coordinate on its evil ends, it still is not normatively authoritative

[9] See Raz, *The Authority of Law,* 17–18, for an official characterization; he then referred to such reasons as protected rather than preemptive.

[10] Raz, *The Morality of Freedom,* 53.

[11] Raz, "Revisiting the Service Conception," 1033–34. Theoretical and political authority are also similar for Raz in being "relational" or piecemeal.

[12] Raz, "Revisiting the Service Conception," 1031.

[13] Ladenson, "Hobbesian Conception of Law."

by the lights of the NJT, which is grounded on the objective moral reasons that apply to citizens.

The existence of functional but highly unjust states is also an object lesson against running the argument the other way and claiming that authority is necessary for coordination. We also have theoretical accounts of how coordination could emerge, even in the state of nature, without the establishment of legitimate authority. In short, the punitive powers of the state (in the proto-form of a gang, perhaps) can be sufficient to stabilize some coordinated practices. It is true that, given the need for coordination in a harsh state of nature, there is some strong reason to obey *any* promising potential leviathan to the extent that obedience helps its chances of success. But this can be captured in first-order terms from the perspective of any given individual, provided some estimation of what other people are doing.[14]

The second argument is even more straightforward: the value of coordination is not piecemeal. The state's coordinative powers benefit everyone. If coordination did much work for authority on its own, we would have a far simpler (and again rather Hobbesian) theory. Instead, it seems entirely clear that coordination powers and epistemic advantages are both necessary conditions for authority: without coordination, the state would be a merely theoretical authority; without theoretical authority, the state is a blunt coercive instrument. So my arguments against treating the state as epistemically authoritative, if they go through, undermine the theoretical structure without requiring any protracted discussion of coordination. Further, the epistemic and coordinative powers of the state are the only real candidates for providing exclusionary reasons that are suitably independent from the "alleged authoritative directive" itself. There are other possible reasons to obey the law, perhaps because it is legitimate in some other sense (e.g., democratically legitimate), but such reasons presuppose the state's authority rather than provide independent rational grounds for it.[15]

It is worth noting another connection between the coordinative role of the state and my arguments against its epistemic authority. Because coordination is a necessary condition on political authority, the state must make laws that are plausible vectors of coordination. This requires, in particular, that the law be relatively *coarse-grained*. A system of laws that attended to every possible circumstance would be cumbersome and impossible to use effectively. How the law applies should in most cases be clear. But simplicity requires the acceptance

[14] Green, *The Authority of the State*, ch. 4, considers similar issues at length. See also Kavka, *Hobbesian Moral and Political Theory*, chs. 4 and 6, for elaboration of coordination emerging from first-order instrumental rationality.

[15] Thanks to an anonymous reviewer for prompting this point.

of error. Coarse-grained law cannot make room for special circumstances that are, nonetheless, relevant to the moral decisions individuals must make.[16] And everyone knows this about the law. This opens up some of the cases to be presented in the next section, in which even nonexperts can have good reason to doubt whether particular commands of the law are justified.

2. Costs of Exclusion

2.1. How Raz Justifies Deference

I will focus mostly on Raz's presentation in *The Morality of Freedom*. Some relevant criticisms about the cogency of exclusionary reasons were lodged soon after that work's publication.[17] But critics have not extended points about the coherence of exclusion to the crucial criticism that if authority is both exclusionary and knowable, we can generate damning counterexamples. That is the task of this section. Further, there has not appeared to be any *alternative* to Raz's account. No matter how troubling the details may be, Raz is surely right that everything cannot be conscious first-order deliberation. Section 3 provides the needed theoretical alternative in order for the criticism to fully land.

The objection is simple: Raz's account commits him to saying that we should obey authorities in some instances in which we clearly should not. I first explain how an unqualified commitment to exclusionary reasoning would generate serious counterexamples and then argue that there is no acceptable Razian way to qualify the account.

Raz's account is highly flexible in that he does not claim authoritative relations obtain generally between the state and citizens. Rather, authority is piecemeal: we must evaluate normal justification at the level of particular agents and particular claims to authority. This feature is also carried over from theoretical authority; coordination-based reasons would seem to fall on everyone equally.[18] We can ask: Given an agent's knowledge, is it in fact rational for them to defer to authority rather than to undertake deliberation themselves? Raz gives the example of the pharmacologist. While there are many laws on which pharmacologists are not experts, they have a great deal of knowledge about drug regulations. So they would not comply with the reasons that apply to them by deferring to the law on questions about which drugs are safe to take;

[16] This is one way of motivating philosophical anarchism: the law by nature cannot be right all the time, so why should we obey when it is wrong? See Simmons, *Moral Principles,* for the classic discussion.

[17] See Gans, "Mandatory Rules"; Edmundson, "Rethinking Exclusionary Reasons."

[18] See again Raz, "Revisiting the Service Conception," 1033–34.

if a pharmacologist is confronted with a decision about whether to take an illegal drug for a rare health condition, it is rational to deliberate directly about the right thing to do.[19] But this is particular to the domain of pharmacology. Pharmacologists ought to defer to the law in other domains. The question is an all-things-considered one about the best procedure for complying with reason in given ranges of cases.

The general picture is something like this. We know we cannot deliberate about every single action. That would not always be the best way to arrive at the correct answers given that our deliberative powers are prone to error, and, besides, it would take up all our time. The proposal is that we can solve these problems by deferring in some ranges of decisions. Just as we often assume that experts know better than we do in our daily lives, we can generally assume (in a decent state) that the law has some epistemic advantage. Even though the law is imperfect, it is generally better than we would do on our own. This does not hold in the particular domains in which we carve out our own expertise, so the law is not authoritative (for us) in those domains. But that cuts down our deliberative burden to a reasonable scope, tailored for each individual epistemic position. This explains why it can be rational to defer; we can know that we are following a procedure with good consequences, even though it requires ignoring the consequences of particular cases.[20]

2.2. Counterexamples

It is harder than it seems to plausibly specify the domains in which we should treat authority as exclusionary. The issue is that the true domain of authority must be sufficiently transparent to function in practical reason. Every agent has to identify which laws to defer to.[21] But once we specify domains of authority in any tractable way, it is clear that the state can make errors *within* its proper (piecemeal) domain. These errors can be significant and transparent enough that deference is perverse. I will give several examples where state errors do not require general expertise to see; each example is meant to illustrate a broad category. I will then consider how Raz's account seeks to avoid such examples. Then I will argue that the resources Raz can deploy to successfully avoid the counterexamples run afoul of the knowability condition.

[19] Raz, *Morality of Freedom*, 74.

[20] It bears repeating that this is not a claim about the full *nature* of authority nor of the moral problem of deference; Raz's account of rational deference is just one part of the view—but it is a necessary part, as I showed in section 1 above.

[21] This "knowability" condition is stated most clearly in Raz, "Revisiting the Service Conception," 1025–26. But it is latent in the main goals of the theory even when unstated: if authority is to be both piecemeal and practical, its contours must be knowable.

The two examples in this section involve, respectively, arbitrary boundaries within the law and internal inconsistencies of authority. A third example, in section 2.3, concerns moral intuitions that the commands of authority are objectionable.

Travel Restriction. The government has reasonably imposed strict travel restrictions on its citizens in response to an ongoing pandemic, splitting its territory into various districts. It is difficult to get a waiver from these restrictions, and it requires a waiting period. All of these features are reasonable on the part of the state, given the perilous circumstances and the risk of exploitation by the selfish if waivers are too easy to procure. Now consider an individual who recognizes the general authority of the state in this domain but who confronts a difficult situation: a loved one in a different district has a serious (unrelated) medical condition and is unable to receive appropriate care under the current conditions. They are suffering. The agent in question is confident in their ability to assist their loved one and can do so with no great personal sacrifice. But doing so requires flouting the state's commands. How should this agent deliberate?

On the unqualified exclusionary reasons view, there is no way to accommodate the powerful intuition that the agent should break the law to aid their loved one. It is not the case that the state has made a clear epistemic error, and even if it had, the person in question has no special expertise on appropriate pandemic travel restrictions. It would not be a good general disposition toward the law to closely evaluate each law to see if there are good personal reasons for disobedience. So the NJT is satisfied, and deference is apparently warranted. But this is seriously counterintuitive.[22]

We can sharpen the case and connect it more clearly to the coarse grain of the law by stipulating that the agent in question lives immediately on one side of the district boundary, while their loved one lives just a few streets over, but on the other side of the boundary. The law has to draw boundaries somewhere. The state cannot serve its coordinative role if it attempts to operate on a case-by-case basis. Everyone knows and accepts this about the law. But it is difficult to take seriously that the law has a decisive epistemic advantage in its decision to place any given person just on one side of the boundary or the other, especially when there is a pressing reason that an individual would prefer to be (or act as if they were) on the other side. This is compatible with the thought that the state has *some* expert reason for placing the boundaries as they did; it just

[22] It may be tempting to reply that if one *really* has good reason to break the law, the service conception simply does not apply. This trivializes authority and should be resisted. See section 2.5 below for an argument, but here I will rely on Raz when he writes that "even legitimate authorities make mistakes. In such cases we should conform with the directive" ("Revisiting the Service Conception," 1023).

does not seem that that reason could pertain to the decision of whether or not to help one's loved one in our case (or in many other structurally similar cases).

Mask Mandate. Famously, the US Centers for Disease Control and Prevention (CDC) was slow to recommend face masks be widely worn in shared spaces during the initial outbreak of COVID-19, instead recommending well into 2020 that masks were only useful for medical providers or for those who knew or strongly suspected that they were infected. This was eventually reversed—but the CDC was then slow to emphasize that masks vary in their efficacy, that surgical masks are preferable to cloth masks, and that (K)N95 masks, in turn, are preferable to standard surgical masks. In each case, it appears that the logic was driven by worries about supply shortages of higher-grade masks. The increase in popularity of cloth masks assuaged the worry that masking *simpliciter* would lead to a supply crisis, but either the worry about supply of higher-grade masks persisted, or the CDC simply did not want to change its guidance again.

Few would deny that the CDC is an extremely strong case of governmental expertise on matters that require a great deal of technical knowledge. The problem is that the CDC was internally inconsistent, most obviously in its reversal on the efficacy of masking for the broad population. And the idea that masks should be preserved for medical providers (who would wear the marks regardless of whether they were infected) but would not be useful for the broader population made no sense to begin with.

This significantly damaged the credibility of the CDC, and the problems of internal logic were clear to non-experts. One *New York Times* op-ed published on March 1, 2020, by Zeynep Tufekci—an academic without medical or biological science credentials—argued that the CDC's official guidance on mask wearing was a mistaken public-messaging strategy, in large part because it was misleading as advice to individuals. This public criticism apparently played a meaningful role in the CDC later changing its official position in April 2020.[23] Tufekci's argument hinged on the points mentioned already: that the CDC policy was inconsistent and that there was an alternative rationale—regarding the supply chain—that made more sense. Once again, the transparency of the error hinged on the state's need to coordinate. The CDC seemingly feared that emphasizing the importance of masks was incompatible with preserving medical supply, even if they had also asked that individuals use masks sparingly until supply could be increased. The means of achieving a desirable coordinated outcome involved damaging their epistemic authority.

So, if one was trying to make a decision about whether to wear masks in general or whether in particular to seek out N95 masks, the CDC in early 2020 was

[23] Smith, "How Zeynep Tufekci Keeps Getting the Big Things Right."

pretty evidently not a good source, despite its unimpeachable epistemic credentials. An immunocompromised person at that stage was better off reasoning on their own and might well have known it. More generally, internal inconsistency is a very important way to identify problems with authority when competing information is simply hard to come by—if a state has sufficient control over the information flow within a society, internal inconsistency may be the *only* way to see through propaganda. But this kind of evidence is accessible to everyone and depends entirely on the particular claims the state makes, not on a general fact about the general domains of expertise that the state and any given individual can claim. (In a propaganda environment, it would make sense to start generally distrusting the state; but, as we have seen dramatically illustrated in the case of COVID-19 vaccines, the failure of the CDC on mask policy was not a good general reason to distrust its advice on other topics.)

In sum: in at least some cases, the justification for treating the law as yielding an exclusionary reason is undermined because it is possible for a citizen of no particular expertise to recognize that the state's commands are particularly fallible in a given case. In other words, deferring to the state does not seem, in such cases, to help the individual comply with the reasons that apply to them better than they could on their own. This leaves two options: we maintain that the state is authoritative in such cases, and subjects should knowingly make mistakes. This looks incompatible with the basic justification of the service conception. More attractively, we can attempt to qualify the service conception to show that this sort of command is not really authoritative. But, because the cases in question do not involve special expertise on the part of subjects, this strategy will need to be even more piecemeal than the standard Razian picture. And I will argue, "robustly" piecemeal authority of this nature cannot satisfy the knowability condition. I consider three possible Razian defenses to this end. The first appeals to emergency circumstances. The second draws a distinction between clear and significant errors. The third attempts to rule bad commands outside the domain of deference.

2.3. *Emergency Exceptions*

The simplest way to qualify Raz's account is to claim that authority does not hold in certain kinds of emergency circumstances. This is a popular move.[24] But it is not clear exactly how it should work. "Emergency" has several connotations. One kind of authoritative emergency is a novel situation for which the

[24] See Raz, *Morality of Freedom*, 46; Tasioulas, "The Legitimacy of International Law," 104; Adams, "In Defense of Exclusionary Reasons," 46n24.

law is unprepared. That kind of case is irrelevant to my counterexamples, which lie in familiar governmental domains.

A second and more intuitive meaning of "emergency" connotes high stakes situations where there is not much time to think. Consider David Estlund's example of the authority of the flight attendant after a plane crash. But this cannot be what Raz means: cases like the plane crash are paradigm cases *for* deference to authority, not cases of exemption from authority.[25]

The third and most relevant kind of authoritative emergency is when a state error is so profound that it immediately delegitimates the authority in and of itself. This does not seem to apply in *Travel Restriction* or *Mask Mandate,* both cases where the fallibility of the law is discernible, but the mistake is not especially profound (in *Travel Restrictions,* the policy itself is not mistaken at all). But emergencies are relevant to another important kind of case.

Moral Crimes. The stakes can go much higher than in my original cases. Consider the conventional and nuclear bombing of civilian populations near the end of World War II; the firebombing campaign in Vietnam; or the killings and maimings of civilian populations as "collateral damage" in Vietnam, Afghanistan, Iraq, or any other "counterinsurgency" campaign. These are all tragic cases; in several cases there appears to have been no remotely plausible just cause for the military operations, so they constitute significant moral crimes. Moreover, this could plausibly be known to some people at the time; we might think, at least, that anyone has good reason not to simply *defer* on the question of the nuclear destruction of entire cities.

But consider the perspective of a bomber pilot. The military, for good reason, has highly deferential norms. Bombing campaigns of massive scale had previously been undertaken, which were at least plausibly justified. And there was a coherent rationale for the late war bombings: that, by their very cruelty, they would end the war sooner and thus save lives in the final balance. This line of reasoning is suspect, and many soldiers might have rejected it. But it cannot be intuitively dismissed the way a nuclear bombing of a neutral city could be. Similarly, one might have gone in for the Domino Theory on which the fate of the world, in some sense, hung on the outcome in Vietnam. We could and should reject these rationales, but it seems plausible that the best general decision procedure for soldiers is quite deferential, and the all-things-considered evaluation of military benefits versus civilian costs is clearly a domain of authority. Nonetheless, it seems that it should be worth deliberating on

[25] Indeed, Estlund is in the business of motivating his account of authority when he gives that example ("Political Authority," 356–58).

participating in a nuclear bombing campaign when the war is largely won. But that is incompatible with unqualified deference to legitimate authority.

So what can we make of the appeal to emergency exceptions? Ultimately, I think, not much: what constitutes an emergency and what reasons obtain within an emergency are questions just as subject to the rationale for deference as the original questions of authority in normal cases. We can see the dilemma played out in miniature with Raz's linkage between emergency circumstances and the possibility that a "directive violates fundamental human rights."[26] While some human rights violations may be completely transparent, other violations involve complex judgments about, e.g., the proportionality of the use of force.[27] Those questions immediately go beyond the epistemic "paygrade" of ordinary soldiers, so we cannot help ourselves to a broad exception for human rights violations, nor emergencies, without undermining the practical function of the service conception. The next section develops a similar line of argument, back in the standard circumstances that do not require any reference to direct intuitions about moral crimes. The dialectic becomes somewhat more complicated, but the conclusion is much the same.

2.4. *Clear and Significant Errors*

One of Raz's central discussions of state fallibility concerns a distinction between clear and significant errors. He recognizes that we should not *stipulate* that significant errors cannot be authoritative; this would require individuals to judge whether any given command is a significant error, which would itself require the expertise we typically lack.[28] Instead, he distinguishes clear mistakes from significant mistakes. Some significant mistakes may be too difficult to detect for personal deliberation to be helpful. And some mistakes, crucially, can be so manifestly clear that they do not require deliberation at all.[29] Raz admits the possibility that a truly horrific state command could be disobeyed on an intuitionistic basis, circumventing deliberation altogether.[30]

[26] Raz, *Morality of Freedom*, 46.

[27] United Nations Human Rights Office of the High Commissioner, *International Legal Protection of Human Rights*, 51.

[28] Raz, *Morality of Freedom,* 47.

[29] In the war examples from section 2.3 above, this would be more like the aggressive invasion of a random state—you might simply *know* such an invasion is wrong, unlike the cases from Vietnam or WWII, which should be immediately *troubling*, but in which one might be brought up short by the Domino Theory or the notion that the nuclear bombings would save lives overall.

[30] "Establishing that something is clearly wrong does not require going through the underlying reasoning. It is not the case that the legitimate power of authorities is generally limited

Raz is correct that some state errors are so transparent that they do not require deliberation about whether the law should be defied. I am happy to grant a non-deliberative proviso for such cases. He is also correct that some state errors are so difficult to identify that the best decision procedure will recommend obedience. But significant state errors do not come in only two varieties—totally opaque or totally obvious. Raz says nothing about the vast middle of this spectrum: significant errors that are partially transparent, or, as I will call them, suspicious. Suspicious state actions are those that are not so obvious that deliberation is otiose but that are troubling enough to prompt an inclination to think or learn more about the matter at hand.

All the cases considered so far can illustrate both transparent and merely suspicious state errors. One version of *Travel Restriction*, mentioned above, might have your needy loved one just down the street if you live near the border. We might intuitively break the law in that case. But in other versions of *Travel Restriction*, aiding your loved one might require traveling some distance, stopping at several gas stations, perhaps a hotel stay. The relevant risks may not be entirely clear, and while the rationale for the placement of the border can *prima facie* be seen to be somewhat arbitrary, it likely is not *completely* arbitrary. There may be mixed messages from public health authorities, which come to the fore in *Mask Mandate*—but depending on the significance and frequency of the inconsistencies, they typically damage institutional credibility rather than eradicate it. The question is how much to trust the institution, given its particular track record. An infinite range of weaker or stronger versions of the cases could be produced. All the argument requires is some range of cases in which the appropriate response is precisely *to deliberate on all the accessible reasons*, including both first-order facts about the command in question *and* second-order facts about institutional credibility. The transparent state error proviso artificially divides the range of possible cases: there are cases in which one should defer and cases in which the state error is so obvious that deliberation is unnecessary. But neither deference nor intuitionistic defiance is attractive in suspicious cases.[31]

One way to put the point is that Raz exaggerates the costs of deliberation. Some salient features are obvious even though not decisive. The cases I

by the condition that it is defeated by significant mistakes which are not clear" (*Morality of Freedom*, 62).

[31] Cf. Perry, "Second-Order Reasons," 933–36, on varying "epistemic thresholds" for ceasing to defer to authority. Perry does not argue that recognizing the mere possibility of error poses a serious problem for Raz. This is because Perry (provisionally) accepts Raz's denial of "partial deference" strategies beyond intuitionism (932), discussed in the text just below. That denial sets up the "all-or-nothing" nature of exclusionary deference.

have developed are ones in which the stakes are clearly high, and the quality of the state's command is—based on what we already know—suspicious. In the *Moral Crimes* cases, the suspicion is based on a strong *prima facie* moral intuition; in the *Travel Restriction* case, it is based on the fact that any boundary-drawing exercise will be partially arbitrary; in the *Mask Mandate* case, it is based on internal inconsistency. An individual undertaking deliberation in such circumstances seems clearly worthwhile.

There is one additional worry we should consider. Raz is concerned that we can fall prey to various personal biases in our deliberation, and these biases might apply equally well to any meta-judgment about whether the state's credibility is undermined in any of the cases mentioned.[32] Raz suggests a promising non-exclusionary strategy that could be used to cope with bias. We might apply a discount rate to our certainty in some cases, taking the authority's reasons to be, e.g., "20 percent stronger than it would otherwise appear to me." Raz dismisses this proposal:

> If, as we are assuming, there is no other relevant information available then we can expect that in the cases in which I endorse the authority's judgment my rate of mistakes declines and equals that of the authority. In the cases in which even now I contradict the authority's judgment the rate of my mistakes remains unchanged, i.e., greater than that of the authority.... Of course sometimes I do have additional information showing that the authority is better than me in some areas and not in others. This may be sufficient to show that it lacks authority over me in those other areas.[33]

This point rests on an odd starting assumption that "there is no relevant information available." It is precisely additional available information that grounds the additional confidence that distinguishes the cases in which our judgment survives the "bias penalty" and those in which it does not.[34] What the bias penalty manifests is the idea that I should not disobey the state on the basis of a deliberation that produces a credence of 0.51 on what the best choice is. We might insist on disobeying only with credences, say, above 0.75. Additional relevant information, such as accidental expertise about my personal circumstances in the context of an arbitrary boundary-drawing law or the state having

[32] Raz, *Morality of Freedom*, 75.

[33] Raz, *Morality of Freedom*, 68–69.

[34] Perry points out that this strategy seems akin to certain familiar cases, e,g., the legal presumption of innocence ("Second-Order Reasons," 933).

been internally inconsistent, is just the sort of thing that can raise one's credence despite the risk of bias and error.

Credence is a subjective evaluation, so there probably is no one general standard for an appropriate "disobedience credence." Such a standard would itself be piecemeal, depending on the stakes of the decision and on the competence and self-awareness of the agent. My deliberation might result in the conclusion that the state *seems* to have made an error but that I am not confident enough to actually disobey. This would result in deference to the state, but not *exclusionary* deference—the decision to defer to the state after an all-in deliberation is a decision, at best, to treat the state *as if* it were authoritative.[35] To treat the state *as* authoritative would have meant restricting deliberation from the start. Of course, as we have seen, treating the state as authoritative is compatible with disobeying in some completely transparent cases—this is akin to setting the appropriate disobedience credence at 1 for all subjects. But that standard is appropriate only for children, if even then; it is not plausible that competent agents do best by restricting their practical reason to solely self-evident state errors.

One further worry could be that the bias is so pernicious that we cannot reasonably apply a bias penalty—we are biased in assessing our own competences and credences, too. That degree of pervasive subconscious bias, however, would presumably also infect the *second-order judgment* distinguishing the domains of our expertise in which authority fails to obtain. If bias is profound, we really would need to turn to a generally less rationalistic account (see section 3 below). If, more plausibly, bias is serious but manageable, then a first-order bias penalty ought to do the trick. The fact that I have extra information that makes me highly confident in this case is good reason to think that this case—but not necessarily this *area*—is one in which I stand a better chance than the authority.[36]

2.5. *Authority's Domain*

Finally, we might press the possibility that authority is really only legitimate when it does not make serious errors. Raz originally handled this thought with the unsatisfactory appeal to clear and significant errors, but he later returned to the thought that there is only legitimate authority over some domain if there is no *part* of the "domain regarding which the person or body can be known to

[35] See Darwall, "Authority and Reasons."

[36] Cf. the classic "rule worship" objection from Smart, "Extreme and Restricted Utilitarianism." Raz seemingly claims, at the limit, that even if I had the word of God that the almost-infallible authority is making a rare mistake in this case, deferring is still my best play.

fail the [epistemic] conditions."[37] The problem again concerns the conditions in which an authority can be known to issue bad commands; it may be that this is another appeal to transparent errors and thus has the same shortcoming as the earlier version.

But there is a stronger available reading of the phrase "can be known." Rather than invoking *ex ante* transparency of errors, it could invoke errors that can be known *after* deliberation or some process of learning. This reading marks a significant change from Raz's original view, on which a command can be authoritative even when one recognizes that it is wrong. But it is consonant with Raz's remark, just prior to the phrase quoted above, that "When the issue is of importance we extend our inquiries and deliberations well beyond what we do when the matter is relatively trifling. The same kind of consideration applies to establishing the existence of authorities."[38]

One worry is the bias concern mentioned at the close of section 2.4: Why are we better positioned to make this second-order judgment about the existence of an authority *qua* action *x* than the first-order judgment about action *x*? But we might set that aside because there is considerable plausibility to the idea that we should proceed relatively undeliberatively with regard to unimportant actions but think carefully about important actions (when we can). That Raz mentions "inquiries" as well as "deliberations" suggests that he is not just concerned with our epistemic state at a given time but also embraces choosing to learn more about a given issue because of its importance.

Presumably, nothing is excluded in this second-order deliberation about whether authority obtains since exclusion follows from authority being known. What will the inquiry consist of? Consider two possibilities. First, one could inquire only about *general* features of the authority relevant to the issue at hand. This path will not avoid the counterexamples; one's inquiries might lead to the conclusion, once again, that the authority is actually very reliable in this domain but just happens to be wrong in the particular case that prompted the deliberation in the first place. So only a stronger possibility helps. We can countenance full-throated deliberation about the case at hand, using whatever we can learn both about the (putative) authority's general features and this specific command.

What does this deliberative picture look like? A special procedure kicks in whenever a command is important. But this is actually too strong because we cannot plausibly inquire about every important law. So some condition of salience will need to be met, which I have called "suspiciousness." In suspicious

[37] Raz, "Revisiting the Service Conception," 1027.

[38] Raz, "Revisiting the Service Conception," 1025.

cases, we recommend unconstrained deliberation. Where has the exclusion gone? It applies, seemingly, only to unimportant cases or to cases in which it never strikes us to deliberate in the first place. In the cases we care most about, nothing is excluded. This is a rather anemic proposal. It should prompt us to ask if there is a more straightforward analysis of the relevant phenomenon. There is. This last reading of Raz—the only one that addresses the problems—is already an account of habituation in all but name.

3. How Habits Help

3.1. Automaticity and Intervention

Despite my criticisms of the exclusionary reason account of deference to authority, Raz is correct that the costs and risks of deliberation are prohibitive in many circumstances. It could be the case that if the choice were between always deliberating and always deferring, it is better to adopt the exclusionary stance. But there is at least one disposition that is better justified on Raz's own terms: habitual obedience. I find the habitual stance appealing, but dialectically it only has to defeat the service conception; there may be additional possibilities.

Habituation has been developed in recent years in other contexts, notably by Pollard and Pettit.[39] The key feature is "intervention control," which characterizes a mental stance toward some routine process for which explicit cognitive attention is not generally necessary, but—crucially—explicit attention can be prompted at any time by unusual circumstances. Pettit gives the example of a cowboy guiding a herd of cattle down a familiar path. Generally, the cowboy simply rides nearby, not actively steering the herd. But if the cattle are spooked, the cowboy should exercise control and restore the herd to the path.[40] A more accessible example is a routine commute between home and work. Most of us do not deliberate on what route we will take on a given day—but if we see, or learn in advance, that there is a construction site in our normal path, we are prompted to deliberate today in particular.

A habit, on my view, is a moderate practical disposition between constant deliberation and principled deference. More precisely, a habit is a trainable, automatic—but defeasible—disposition to act in a certain way in a certain range of circumstances ("the usual"). Let us say that a habit is justified the same way Raz tells us authority is justified: if and only if relying on the habit is generally the best way to conform to reasons that apply to us.

[39] Pollard, "Can Virtuous Actions Be Both Habitual and Rational?"; Pettit, "The Inescapability of Consequentialism."

[40] Pettit, "Inescapability," 45–46.

Let me say a word about each of the noted features of habits. A habit is trainable. An unalterable *instinct* is not a tool available in practical reason, but a habit is. This training might be purposeful or might simply crop up with sufficient repetition.[41] A habit is automatic. As opposed to deliberation, which for Raz is seemingly transparent to the agent even when exclusionary rules are followed, following a habit drops below one's awareness.[42] But this subconscious automaticity is defeasible in the sense of intervention control—as with the cowboy, unusual circumstances prompt unusual deliberation.

Consider an example of developing a skill. When learning to play tennis, much of what it *means* to develop skill is for more and more patterns of movement and behavior to drop into automatic background processes.[43] Consciously deliberating on each shot is "playing tight" and leads to poor results. There are some advantages—perhaps better tactics—in deliberating on each shot. But it will fail in terms of the overall goal of winning the match. This is true of skill development generally. We can almost always perform better by relying on automatic processing. (Of course, not *fully* automatic; if our opponent is injured, intervention control kicks in to stop us from smashing the next ball at them.)

This translates reasonably directly to dispositions toward the law. My habit of following traffic laws both improves my performance—my reaction time is better when deliberative processing is not involved—and avoids some incorrect judgments that I should break the law in mundane circumstances. This morally justifies taking up the right kind of habits to the right degree. But when circumstances are genuinely unusual and there is time to invoke intervention control, such as when I need to flout traffic laws to take someone to the hospital, the habit is set aside. The counterexamples developed in section 2.2 are clear cases where intervention is warranted—even if in some other cases time is too short or information is too lacking.

3.2. *The Superiority of Habits*

There is admittedly something unsettling about the role of automaticity. If our topic is *normative* powers of authority, should we not comply knowingly? But recall that the focus here, as in Raz's NJT, is how authority can be justified. Authority provides a benefit—thus the *service* conception. I agree with Raz

[41] Thus, my sense of habits collapses Owens's distinction with consciously chosen personal policies ("Habitual Agency," 99–100). This is just a terminological simplification.

[42] See Snow, *Virtue as Social Intelligence*, for helpful conceptual and empirical discussion of automaticity. See also Arpaly and Schroeder, *In Praise of Desire*, sec. 2.7.

[43] In Kahneman's terms, intuitive "system 1" processing rather than deliberate "system 2" thinking (*Thinking, Fast and Slow*).

that a justificatory account, rather than a mere conceptual analysis, is what we should want. But if the NJT does the work, then it is fair to argue against the exclusionary analysis by proposing a *better-justified* disposition toward the law. When you can get a better deal, you switch services.

One might press this further and ask whether a habitual account can be an account of obeying authority at all. As R. P. Wolff says, "[obedience] is a matter of doing what he tells you to do *because he tells you to do it.*"[44] My position is that it is neither here nor there whether the habitual account meets such a conceptual criteria; the point of dialectical importance is what the best deliberative stance toward putative political authorities is, given Raz's own (defensible) standard of justification. One possible conclusion is that Raz's service conception cannot consistently be a service conception *of authority* and should instead be read as skeptical of authority. We nonetheless can defend a service *justification of the state* based on habituation. I prefer to leave conceptual space for a deflationary account of authority, on which authority is not quite what we might have thought but still warrants the title. But the substantive conclusion about justification is the central point, not the conceptual question.[45]

I have indicated two arguments for the superior justification of habits. First, the habitual account—making use of intervention control—avoids the counterexamples to the exclusionary account. Second, relying on automatic choice procedures is a normal element of becoming skilled in any domain, and there is no obvious reason that competence at navigating the law should be different. This section illustrates an additional theoretical advantage: the habitual account improves on an awkward distinction Raz draws between practical deliberation and mere consideration or reasons. The principal concern is to avoid acts that are grounded on excluded reasons. But we are free to *consider* excluded reasons—"So long as one knows that one's reflections will not affect one's actions." John can think about whatever he likes but "is only acting correctly if he disregards the excluded reasons in his deliberation."[46]

This opens the possibility of considering a case closely enough that it becomes clear that the excluded reasons actually should be decisive. One cannot know in advance how reflection will go. Part of the point of idle contemplation is that it sometimes leads to action. More pointedly, there is always a chance that idle reflection on some generally good rule will yield continued general endorsement of the rule but some particular conclusion about making an exception.

[44] Wolff, *In Defense of Anarchism*, 9.

[45] Thanks to an anonymous reviewer for pressing this point.

[46] Raz, *Practical Reason and Norms*, 184–85.

Consider a case where the basis for an exclusionary (or habitual) disposition is the cognitive burden of constant deliberation. Say I am permitted to go home early from work if my day's tasks are done, but doing this excessively is frowned upon. I only finish work early occasionally, so I decide simply generally not to deliberate on whether I have good reason and sufficient political capital to knock off early. One day at lunch, I idly contemplate the possibility of taking off early in the course of a conversation about the cogency of the general policy about leaving work on time. I realize that *today*, which I was merely using as an example in conversation, is actually an exceptionally good time to leave early. Due to my general policy, I have not left early in months, and I have nothing at all useful to do. If I treat my (well-justified) rule as exclusionary, however, I must maintain the firm wall between idle contemplation and practical deliberation. So I should not take off work early today because it would be too costly to deliberate about such cases generally, even though I have *already deliberated about this case*. This is a bizarre result that intervention control naturally avoids.

3.3. *Habit Formation*

Habit formation should play an important role in practical reason. Good habits are very valuable; they cope with our cognitive limitations without causing too many errors. That makes developing good habits a relevant part of first-order reasoning about what to do. Habits are *trained* automatic dispositions; whenever one acts in accordance with a habit, it is trained further, and when one violates a habit the training is undermined. A habit's weakness or strength can be thought of as how reliably deliberation is circumvented or truncated in the relevant range of circumstances. How we act now affects our choice procedures in the future.

Given the value of good habits, maintaining a habit can itself be a reason to act in accordance with the habit. This partially recaptures the spirit of the exclusionary account. Exclusion is typically tightly linked with content-independence. Standard examples of content-independence are reasons to do what someone says, regardless of what in particular they say; recall Raz's example of following a friend's advice in order not to offend them. Adams describes this as a reason due to the source or "container" of a specific act.[47] We might think of a habit as a container for an action; the action has whatever first-order merits and demerits but has an additional reason in its (dis)favor in virtue of maintaining or undermining a habit. The weight of this reason will vary with many factors;

[47] Adams, "In Defense of Content-Independence," 147.

presumably not every violation of a habit is equally meaningful, and particular individuals may tend to form habits that are more or less fragile.

Now, one might reserve content-independence more strictly for reasons due directly to the *standing* of the source or container. Habits do not have standing in that sense but instead provide an indirect rationale for obeying (some) commands, whatever they may be.[48] But habituation fits the intuitive way of explaining content-independence and shows why we should sometimes obey even a poorly justified command. Of course, maintaining a habit is only so valuable. While habit formation and maintenance partially captures the appeal of exclusionary reasons, it does not expose the habitual obedience account to the weightier counterexamples raised against the service conception of authority.[49]

This may suggest a line of orderly retreat for the service conception. Habituation offers an attenuated version of content-independence; but what about a revised analysis of authority that says that an authoritative command directly provides a content-independent, but not exclusionary, reason? The problem with this proposal is that the service authority provides is precisely to *settle* practical deliberation. Exclusionary rules are decisive, which in turn motivates the piecemeal account of authority—decisive authority is only a benefit if the authority will generally decide better. A retreat to content-independence without exclusion unravels the distinctive Razian story. Without exclusion, the law is not decisive; if the law is not decisive, it is not clear why we should say that authority is piecemeal. We might then say that authority yields general, defeasible, content-independent reasons to obey the law. This is precisely the *traditional* analysis of political obligation, attacked most famously by John Simmons.[50] This analysis retains many defenders. But it is not the Razian analysis.

[48] It is easy to slide between content independence residing in the standing of the reason giver versus the neutrality of the reason across particular actions. Adams describes an advice-style case ("In Defense of Content-Independence," 158–59), where what is really at stake is the effects on a relationship as content independent. But in discussion of threats, he says, with Raz, that penalties (and presumably downstream causal effects generally) are actually part of the content of a threat, which Raz considers merely a content-independent reason to *believe* rather than to act (Adams, "In Defense of Content-Independence," 156; Raz, *Morality of Freedom*, 36). But the threat case seems structurally similar to the advice case. The intuitive phenomenon that embraces deontological authority, threats, habituation, and concern for relationships might be better termed content *neutrality.*

[49] Some readers will have been reminded of Darwall's distinction between directives being treated as authoritative and directives actually being authoritative ("Authority and Reasons"). Another way of putting the point of the above paragraph is that habituation stays on the "treating as if authoritative" side of that distinction—and even the reasons to "treat as if" have limits.

[50] Simmons, *Moral Principles*. Raz discusses political obligation, which for him is always distinct from authority, in *The Authority of Law*, ch. 12, *The Morality of Freedom*, ch. 4,

The habitual account points out that the state can provide benefits merely because the law is a salient anchor for a habit of obedience. This provides an indirect general reason to obey the law in order to maintain the beneficial habit. The weight of that reason will vary with the justice of the state and the elasticity of a given individual's habit. But the automatic nature of habits means that they very often *will* settle deliberation—indeed, conscious deliberation will never get started. Again, habituation explains some intuitive features of authority while avoiding unattractive results.

Habits are not always beneficial. Some habits are bad—patterns of behavior that one follows unthinkingly but that, in fact, yield worse results than direct deliberation or some alternative habit. Just as there is a general reason to form or maintain good habits, there is a general reason to break bad habits. The worry about the slippery slope from disobedience into anarchy is only one side of the coin. Any habit, surely including the habit of obeying the law, can become overly entrenched and thus act as a false principle, so we must take care in the other direction as well. The ideal is equipoise, recognizing slippery slopes on both sides.

The next two sections address objections: first, that habits themselves may be analyzed as exclusionary; second, that habits may fail to stabilize political institutions in the face of collective action problems.

3.4. Habits and Exclusion

Here is a challenge. Habits (and policies) are sometimes themselves discussed as having an exclusionary character in practical deliberation.[51] Have I replaced one exclusionary notion with another? No. Where theorists of habits invoke exclusionary considerations, they either do not or should not mean what Raz means. The shared insight is that some dispositions (habits, policies, principles, plans) serve to prevent (re)consideration of choices in some range of circumstances. But this range is not well characterized by excluding certain types of reasons as practically irrelevant. This is easy to miss. Owens writes, of a disposition to always go on a daily run, that "your policy has an exclusion zone around it, one that rules out consideration of discomfort but not of threats to your health."[52] But this cannot be correct. It is true that some discomfort will not prompt deliberation. A chilly day might be regrettable, but it is not relevant

esp. sec. 4, and "Revisiting the Service Conception," 1004–12. See Dagger and Lefkowitz, "Political Obligation," for general discussion. The habitual obedience account seems to me compatible with philosophical anarchism, but I do not think it requires it.

[51] Owens, "Habitual Agency," 105. See also Holton, *Willing, Wanting, Waiting*, 101–5; Bratman, *Intention, Plans, and Practical Reason*, ch. 5.

[52] Owens, "Habitual Agency," 100.

to my habit. However, a freezing cold day with hail is relevant to any normal running habit, even if it is not a threat to my health and does not pose any other kind of emergency.

Instead of focusing on reasons that are categorically excluded from deliberation, habits should be understood in terms of intervention control. In normal circumstances, we rely on our trained disposition rather than deliberation, and in that sense, many possible considerations are excluded. But any sufficiently surprising circumstances can prompt deliberative intervention. Once we are jarred into deliberation, we undertake an all-things-considered deliberation in which no reasons are excluded. Which considerations prompt deliberative intervention depends on what considerations are evident to us, which is quite contingent. I should not seek out all the possible construction sites on my commute every day, but, as illustrated above, I should not on that basis ignore the construction site I *already* know about.

The answer to the objection, then, is that habits are not exclusionary in the same way as exclusionary reasons. An exclusionary reason is a reason that is deemed irrelevant *within* an ongoing, conscious deliberation. A habit is a disposition not to deliberate at all under a range of circumstances. Once that automatic pattern is disrupted, deliberation proceeds unimpeded.[53]

3.5. *Stabilizing Institutions*

Is habitual obedience enough to do what the exclusionary reasons account sets out to do—namely, explain good practices of epistemic deference and stabilize coordination goods? Plausibly, yes. There is little question that most people will develop a habit of obedience to the law in reasonably just societies. Respect for the law is part of many cultures and encouraged by parents and other influences. In a good society, it will often be natural and convenient to do what the law says, so the overall disposition will be further buttressed. And, of course, fear of punishment is always available as a general reason to obey. This seems sufficient for coordination goods of the kind Raz emphasizes.[54] Given a general habit of obedience, whatever the law says will be salient, such that in relatively neutral cases of coordinating conventions—such as which side of the road to drive on—coordination will be easily achieved. And the benefits of coordination goods will further ensconce routine obedience to the law.[55]

[53] Thanks to an anonymous referee for requesting clarification here.

[54] Raz, *Morality of Freedom*, 48–52.

[55] See Buchanan, "Institutional Legitimacy," 64–65, for a related discussion of what he calls the "virtue of law-abidingness." There is also an affinity with Austin's command theory of law (*The Province of Jurisprudence Determined*), with punishment acting to stabilize habits

All this is morally valuable to the extent that habits are justified. Just and reasonable governance will have a positive feedback loop with robust habits; the better the government, the less cause there will be to exert intervention control, causing habits to become more stable, which in turn allows the state to operate more smoothly and sympathetically because the more habitually citizens obey, the less the state must be concerned with punitive enforcement of the law.

Epistemically, one might worry that because habits do not focus as tightly on the epistemic advantages possessed by the law, the habitual stance will be a harmfully less deferential stance when the law truly is epistemically advantaged. If there is a generally established habit of obedience, the difference between habituation and exclusion will only appear in suspicious cases. In such cases, habitual obedience does entail extra deliberation compared to the exclusionary reasons account, and this may come at some cognitive cost. But there is no reason for epistemic modesty to disappear altogether. If my habit is brought up short by a surprising circumstance, but my deliberation can hardly proceed because I do not know enough, then epistemic deference is perfectly appropriate. This added deliberative step seems a small price for the moral benefit of recognizing when the law is performing quite badly in ways that *are* epistemically accessible for a given agent.

4. CONCLUSION

Raz's theory of practical authority begins with the move from what actions are normally justified to what disposition toward authority is generally justified. There is more to his account of political authority, but this move undergirds that account and by itself sets up the highly influential notion of exclusionary reasons. I have argued against this central justificatory move. Many cases are not normal, and the best-justified disposition is the one that does best across *all* cases, not in a subset—no matter how familiar. One might draw a parallel with act-utilitarian critiques of rule-utilitarianism. Just because a rule fares best *among rules* does not itself explain why any act falling under that rule is substantively correct. The act-utilitarian then faces a profound challenge because we do need *some* tractable decision procedure. But, regarding authority, I have provided—while not quite a conscious decision procedure—an attainable *stance* in practical reason, which I have argued fares better than the exclusionary stance. If the habitual stance is indeed better justified than the exclusionary stance, we have a better way to navigate our perplexing epistemic world.

of obedience. But the main aims of my argument do not concern the concept of law, so I will not pursue the connection.

Exclusionary reasons are unnecessary—and so the service conception is cut off at the knees.[56]

University of Arizona, Philosophy
travis.c.quigley@gmail.com

References

Adams, N. P. "In Defense of Content-Independence." *Legal Theory* 23, no. 3 (September 2017): 143–67.

———. "In Defense of Exclusionary Reasons." *Philosophical Studies* 178, no. 1 (January 2021): 235–53.

Austin, John. *The Province of Jurisprudence Determined.* Edited by Wilfrid E. Rumble. Cambridge: Cambridge University Press, 1995.

Bratman, Michael E. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.

Buchanan, Allen. "Institutional Legitimacy." In *Oxford Studies in Political Philosophy*, vol 4, edited by David Sobel, Peter Vallentyne, and Steven Wall, 53–78. Oxford: Oxford University Press, 2018.

Dagger, Richard, and David Lefkowitz. "Political Obligation." *Stanford Encyclopedia of Philosophy*, (Summer 2021). https://plato.stanford.edu/archives/sum2021/entries/political-obligation/.

Darwall, Stephen. "Authority and Reasons: Exclusionary and Second-Personal." *Ethics* 120, no. 2 (January 2010): 257–78.

Edmundson, William A. "Rethinking Exclusionary Reasons: A Second Edition of Joseph Raz's *Practical Reason and Norms*." *Law and Philosophy* 12, no. 3 (August 1993): 329–43.

Enoch, David. "Authority and Reason-Giving." *Philosophy and Phenomenological Research* 89, no. 2 (September 2014): 296–332.

Estlund, David. "Political Authority and the Tyranny of Non-consent." *Philosophical Issues* 15, no. 1 (October 2005): 351–67.

Gans, Chaim. "Mandatory Rules and Exclusionary Reasons." *Philosophia* 15, no. 4 (January 1986): 373–94.

Green, Leslie. *The Authority of the State*. Oxford: Clarendon Press, 1988.

Hershovitz, Scott. "The Role of Authority." *Philosophers' Imprint* 11, no. 7

[56] Thanks to Allen Buchanan, Tom Christiano, Max Kramer, Michael McKenna, and Dave Schmidtz for comments, discussion, or both. Particular thanks to Andrew Lichter and especially Steve Wall for ongoing discussion and suggestions throughout the writing of this paper.

(March 2011): 1–19.

Holton, Richard. *Willing, Wanting, Waiting*. Oxford: Oxford University Press, 2009.

Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

Kavka, Gregory S. *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press, 1986.

Ladenson, Robert. "In Defense of a Hobbesian Conception of Law." *Philosophy and Public Affairs* 9, no. 2 (Winter 1980): 134–59.

Owens, David. "Habitual Agency." *Philosophical Explorations* 20, no. s2 (October 2017): 93–108.

Perry, Stephen R., "Second-Order Reasons, Uncertainty and Legal Theory." *Southern California Law Review* 62 (1989): 913–94.

Pettit, Philip. "The Inescapability of Consequentialism." In *Luck, Value, and Commitment: Themes From the Ethics of Bernard Williams*, edited by Ulrike Heuer and Gerald Lang, 41–70. Oxford: Oxford University Press, 2012.

Pollard, Bill. "Can Virtuous Actions Be Both Habitual and Rational?" *Ethical Theory and Moral Practice* 6, no. 4 (December 2003): 411–25.

Railton, Peter. "Practical Competence and Fluent Agency." In *Reasons for Action*, edited by David Sobel and Steven Wall, 81–115. Cambridge: Cambridge University Press, 2009.

Raz, Joseph. *The Authority of Law: Essays on Law and Morality*. Oxford: Oxford University Press, 1979.

———. *The Morality of Freedom*. Oxford: Oxford University Press, 1988.

———. *Practical Reason and Norms*. Oxford: Oxford University Press, 1999.

———. "The Problem of Authority: Revisiting the Service Conception." *Minnesota Law Review* 90, no. 4 (2006): 1003–44.

Simmons, A. John. *Moral Principles and Political Obligations*. Princeton: Princeton University Press, 1980.

Smart, J. J. C. "Extreme and Restricted Utilitarianism." *Philosophical Quarterly* 6, no. 25 (October 1956): 344–54.

Smith, Ben. "How Zeynep Tufekci Keeps Getting the Big Things Right." *The New York Times*, August 23, 2020. https://www.nytimes.com/2020/08/23/business/media/how-zeynep-tufekci-keeps-getting-the-big-things-right.html.

Snow, Nancy E. *Virtue as Social Intelligence: An Empirically Grounded Theory*. Oxford: Routledge, 2009.

Tasioulas, John. "The Legitimacy of International Law." In *The Philosophy of International Law*, edited by Samantha Besson and John Tasioulas, 97–118. Oxford: Oxford University Press, 2010.

United Nations Office of the High Commissioner for Human Rights.

International Legal Protection of Human Rights in Armed Conflict. New York: United Nations, 2011.

Wolff, Robert Paul. *In Defense of Anarchism*. Berkeley: University of California Press, 1998.

Journal of Ethics and Social Philosophy
VOL. 26, NO. 1 · NOVEMBER 2023

https://doi.org/10.26556/jesp.v26i1.1692

RADICAL COGNITIVISM ABOUT PRACTICAL REASON

William Ratoff

Might practical reason be a species of theoretical reason? Can we make sense of practical deliberation as a special kind of theoretical calculation? Common sense teaches us that the practical and theoretical aspects of thought are quite different in nature—whereas practical reason concerns itself with *what to do*, theoretical reason is concerned rather with *what to believe*. The differences between their natures come in two broad kinds: psychological and normative. First, the constitutive psychological ingredients of these two species of thought differ. Consider, for instance, an episode of practical deliberation. It might take desires, beliefs, and intentions as input. And it issues in (further) intentions, or else in adjustments to the inputted intentions, as output. In contrast, theoretical reasoning takes only cognitive states (such as beliefs and credences) as input and produces only further cognitive states, or else adjustments to the inputted cognitive states, as output.

Second, the norms governing these two modes of thought diverge. On the face of it, theoretical reasoning is subject to epistemic norms alone.[1] Epistemic norms include requirements of theoretical rationality, such as the prohibition against believing contradictory propositions, and considerations that count as evidence in favor of believing one proposition or another. In contrast, practical deliberation is governed (in addition) by distinctively practical norms—that is, by requirements of practical rationality, such as the prohibition against intending to perform incompatible actions, and by practical reasons that count in favor of acting in this way or that.

In light of these manifest differences in natures, the prospects for any proposed reduction of the faculty of practical reason to a faculty of theoretical reason may look bleak. After all, it seems like it is one thing to be weighing up what you should (intend to) do, in light of your various reasons for action, and quite another thing altogether to be figuring out what you should believe,

[1] Adler, *Belief's Own Ethics*; Shah, "A New Argument for Evidentialism"; Parfit, *On What Matters*, vol. 2; Way, "Two Arguments for Evidentialism."

in light of your evidence. Nevertheless, in this paper, I pursue this project of seeking to reduce practical reason to a species of theoretical reason. Since these faculties are both psychological and normative in nature, the proposed reduction precipitates a total reduction of the practical—attitudes, reasoning, and norms—to the theoretical. In particular, practical attitudes—intentions and desires—are reduced to beliefs; practical reasoning is reduced to a kind of theoretical reasoning; and practical normativity is reduced to a variety of epistemic normativity. In my terms, it entails a "radical cognitivism about practical reason."

This picture of the mind will likely appear highly revisionary. On the precisified model to be defended, an agent engaged in practical deliberation—that is, deciding *what to do*, given her (believed) reasons for action, or her desires and means-end beliefs—is really just trying to predict what *she is going to do*, given the evidence available to her. Hence, an agent's intentions to act must really be certain of her beliefs concerning what she is going to do, and her reasons for action are revealed to be a species of her reasons to believe that she will so act. Even her desires are reduced to cognitive states. The mind ultimately consists purely in cognitive states that are governed solely by epistemic norms.

However, in a certain sense, this picture is not really revisionary at all. Radical cognitivism about practical reason, as I conceive it, is *not* an eliminativist account of the practical aspects of reality. It is *not* saying that the practical attitudes do not exist, and that the mind is rather just a mass of cognitive states. Nor is it saying that practical norms or practical reason itself are unreal, and that only epistemic norms and theoretical reason exist. No—all it is saying is that the practical aspects of reality *reduce* to the theoretical aspects. And such a view is straightforwardly inconsistent with the nonexistence of the practical.

Why be interested in this theory of practical reason as a species of theoretical reason? What are its virtues? In short: parsimony, both psychological and normative. After all, why posit two fundamental modes of reason—one theoretical in nature, the other practical—when we can make do with just one? First, this reduction unifies and streamlines our theory of the mind: it promises to explain behavior through appeal to just *one kind* of mental state, theoretical attitudes, playing by *one set* of psychological rules, rather than by reference to a *plurality* of such states playing by *different sets* of rules. Second, our normative theory is likewise unified and economized with no loss of explanatory power: normative reality is held to bottom out in epistemic norms alone, entities already posited by our normative theory.[2] Third, this theory of practical reason, I claim, vindicates (limited) forms of moral rationalism and prudential rationalism—the

[2] Moderate cognitivism about practical reason can be motivated through appeal to considerations of normative parsimony: for example, by citing the fact that it allows us to explain certain requirements of practical rationality in terms of certain *already posited*

doctrines, respectively, that we have (some) reasons to be moral and prudent that are independent of our desires.[3] And it does this in a novel way—quite unlike Kantian or realist strategies for defending these conclusions.[4]

Of course, these virtues are contingent upon the proposed reduction of the practical to the theoretical being successfully executed. After all, if we cannot make sense of practical thought and motivation as a species of theoretical mentation, or if we cannot preserve common sense concerning what we have reason to do—enough to be sensibly endorsed, at least—under the new regime, then radical cognitivism about practical reason will fall at the first hurdle: namely, that of accounting for the (behavioral and normative) data. Given this, my aim in this paper is simply to begin this task of showing how we might make sense of practical thought as a species of theoretical cogitation.

I cannot hope to address, in one paper, all aspects of this radical cognitivism about practical reason. Distinct aspects of the radical cognitivist's project—such as her theory of the mind or theory of practical normativity—demand individual attention. Hence, in this paper I will restrict myself to just investigating the cognitivist theory of intention and means-end practical reasoning

requirements of theoretical rationality. In other words, that such a picture allows us to unify and economize our normative theory.

[3] Schafer-Landau, *Moral Realism*.

[4] Korsgaard, *Sources of Normativity*. However, I cannot adequately discuss this third virtue in this paper, since it concerns an aspect of the radical cognitivist's reduction—her theory of practical norms—that I lack the space to introduce here, and which I rather develop and defend in a separate paper (Ratoff, "Practical Reason as Theoretical Reason"). Another reason to pay attention to radical cognitivism about practical reason is that it *may* be an entailment of the prediction-error minimization (PEM) model of the mind that is currently ascendant in cognitive science (Friston, "A Theory of Cortical Responses"; Friston, Kilner, and Harrison, "A Free Energy Principle for the Brain"; Friston, Adams, and Montague, "What Is Value"). This theory of the mind has recently received a lot of attention from philosophers of cognitive science (Hohwy, *The Predictive Mind*; Clark, *Surfing Uncertainty*). According to this theory, all the mind ever fundamentally does is make hypotheses about the environment, generate prediction errors by comparing its predictions with its sensory data, and use these prediction errors to update its representation of the world (Friston, Kilner, and Harrison, "A Free Energy Principle for the Brain"; Clark, "Whatever Next?"; Hohwy, *The Predictive Mind*). (A prediction error is the difference between some prediction and the corresponding observation.) On the face of it, PEM entails that, fundamentally, all mental states are cognitive states, all practical reasoning is theoretical reasoning, and all practical norms are really epistemic norms. Indeed, as one of the principal proponents of PEM, the neuroscientist Karl Friston, puts it, this picture entails that "value is evidence" (Friston, Adams, and Montague, "What Is Value"). Critically, however, not all those working on PEM agree that it entails a wholesale reduction of the practical to the theoretical (cf. Clark, "Beyond Desire?"). Still, taken at face value, PEM looks to entail radical cognitivism about practical reason. This, I think, gives us another reason to take it seriously.

to which the radical cognitivist is committed.[5] On this psychology, your intentions to act are predictions about what you are going to do, formed in light of evidence alone, and means-end practical reasoning is a variety of theoretical inference concerning the likely causes of your predicted future actions. My discussion will center on developing and critically examining the radical cognitivist's options for satisfying the desiderata of any adequate theory of intention and means-end practical reasoning—for example, whether she can explain how mere beliefs can occupy the functional role of intention, or accommodate the commonsense distinction between intending to do something and merely foreseeing that you will do it, purely through appeal to cognitive states and the epistemic norms governing them. Unlike other cognitivists about practical reason, the radical cognitivist reduces all practical reasoning to theoretical reasoning and all practical norms to epistemic norms. She therefore faces *unique* challenges in accounting for the basic desiderata of any adequate theory of intention and means-end practical reasoning: whereas other cognitivists can appeal to *sui generis* practical states (desires) and norms, the radical cognitivist is restricted to the sparse resources—cognitive states and the epistemic norms that govern them—to which she has restricted herself. [6]

Moderate cognitivism about practical reason has been defended now by a plurality of philosophers—including but not limited to David Velleman, Jay Wallace, Kieran Setiya, and Jacob Ross.[7] Such moderate cognitivists hold that certain aspects of practical reason are really instances of theoretical reason. For example, such cognitivists hold that intentions are, or involve, beliefs and that certain norms of practical rationality just are, or can be explained in terms of, certain norms of theoretical rationality. This project can be motivated through appeal, among other things, to considerations of normative parsimony: Why posit *sui generis* practical norms when we can make sense of them as a species of epistemic norm already posited by our normative theory? Radical cognitivism about practical reason, then, is simply the souped-up version of this project taken to its ultimate limit.

[5] Consequently, I will address topics such as the radical cognitivist's theory of normative judgment and whether she can accommodate the possibility of akrasia, etc., not in this paper but rather in a separate paper ("Theoretical Reason as Practical Reason") that concerns the radical cognitivist's theory of practical norms, since these topics presuppose acquaintance with said theory of practical norms.

[6] Harman, "Practical Reasoning"; Setiya, "Practical Knowledge."

[7] Velleman, "Practical Reflection" and *Practical Reflection*; Wallace, "Normativity, Commitment, and Instrumental Reason"; Setiya, "Practical Knowledge"; Ross, "How to Be a Cognitivist about Practical Reason."

This sets the agenda for this paper. My strategy will be to address only those aspects of the radical cognitivist's theory that are distinctive and where she faces unique challenges, and to treat more briefly, or even bracket, those elements that are shared with other cognitivist theories of practical reason. Cognitivism about intention and practical reason is a prominent view in the literature—defended, in various forms, by Paul Grice, Robert Audi, Gilbert Harman, Wayne Davis, David Velleman, Jay Wallace, Kieran Setiya, Jacob Ross, and Berislav Marušić and John Schwenkler—and I do not want to simply rehash any well-trodden dialectical ground.[8] Rather, I will just assume that these more modest varieties of cognitivism are defensible, an assumption that will allow me to avoid relitigating here a number of disputes.[9]

Before we continue, I should briefly address the comparison between radical cognitivism about practical reason and David Velleman's theory of practical reason.[10] The parallels here cannot be ignored: on both models, intentions to act are identified with certain predictions about what you are going to do. And, in both pictures, you are moved—in virtue of your nature as an agent—to act in the way that you rationally expect yourself to act. Consequently, on both views, evidence concerning your future actions can be apt to constitute a reason for you to act in those ways. The similarities, however, end there. Velleman's cognitivism about practical reason is not one of complete reduction of the practical to the theoretical: desires in his picture are left as *sui generis* practical states. Nor, most critically, are practical norms reduced *en masse* to a species of epistemic norm: there are reasons, on Velleman's view, for you to act that do *not* reduce to some kind of evidence about what you will do—for example, a *sui generis* practical reason to *F* given by your desire to *F*. (The radical cognitivist, of course, denies this.) Nevertheless, the deep similarities just cataloged make

[8] Grice, "Intention and Uncertainty"; Audi, "Intending"; Harman, "Practical Reasoning"; Davis, "A Causal Theory of Intending"; Velleman, "Practical Reflection" and *Practical Reflection*; Wallace, "Normativity, Commitment, and Instrumental Reason"; Ross, "How to Be a Cognitivist about Practical Reason"; Marušić and Schwenkler, "Intending Is Believing."

[9] So, for example, to avoid simply rehearsing defensive moves that have already been made in the literature, I will not discuss the matter of whether or not the cognitivist thesis that intending to *F* entails believing that one will *F* is tenable in light of various counterexamples suggesting that there are circumstances in which one can rationally intend to *F* but cannot rationally believe that one will *F*. Rather, I will simply assume that this doctrine is defensible. For a defense of this cognitivist thesis, the interested reader can consult Harman, "Practical Reasoning"; Ross, "How to Be a Cognitivist about Practical Reason"; or Marušić and Schwenkler, "Intending Is Believing"—although for a recent critique, see, for example, Paul, "How We Know What We're Doing" and "Intention, Belief, and Wishful Thinking."

[10] Velleman, "Practical Reflection" and *Practical Reflection*.

Velleman's theory the natural reference point for any discussion of radical cognitivism about practical reason.

The structure of the rest of this paper goes like this. In section 1, I outline and develop the precisified version of radical cognitivism about practical reason to be defended here. In section 2, I show how the radical cognitivist can generate an adequate theory of intention and means-end practical reasoning simply by endorsing the standard view in the philosophical literature concerning the propositional content of intentions. Such a theoretical move allows the radical cognitivist, I claim, to explain how mere beliefs could occupy the functional role of intention—to accommodate the commonsense distinction between intending to do something and merely foreseeing that you will do it and to account for the distinction between our telic intentions and our instrumental intentions, purely through appeal to the sparse resources to which she has limited herself. Last, in section 3, I show how the radical cognitivist can account for the "forward-looking" orientation of practical reason—in particular, the fact that no rational agent ever intends to perform some action without taking it to promote their ends.

1. Radical Cognitivism about Practical Reason

How could practical reason be a branch of theoretical reason? How can we make sense of motivation and practical deliberation with such sparse resources—as a species of cognition and theoretical calculation? Consider some arbitrary episode of practical reasoning. Suppose that I find myself in a novel situation and ask myself, "What shall I do next?" I consider some of the various actions that I could now perform in light of their likely upshots. I then find myself, as a result of this process, concluding that I will perform one of these actions.

Now, this is supposed to be a description of practical reasoning. But nothing that is described here seems *exclusively* practical. There is nothing described that rules out the hypothesis that this is in fact an episode of theoretical reasoning. After all, theoretical reasoning can result in conclusions about what I will do. For example, I can confidently predict now, in light of abundant evidence, that I will one day retire and try to enjoy my remaining days on this earth in a more relaxed fashion. And such theoretical conclusions about what I will do can be formed purely in light of reflection on their likely upshots. For example, I might have concluded that I will one day retire, not because I know that people like me standardly retire at some point, but rather in light of my evidence that retirement standardly produces more opportunities for leisure and that I tend to do things that are likely to produce opportunities for leisure.

This, in essence, is how I propose that practical reason might turn out to be a variety of theoretical reason. Practical reasoning, on the advertised view, commences with you attending to the various outcomes that you could bring about. Since you have good evidence—formed in light of a lifetime of experience—that you will act to bring about certain outcomes as *ends,* for the sake of no further purpose, you will (if rational) be moved to form beliefs that you will so act (to bring about said outcomes as ends). These beliefs constitute your *telic* intentions—that is, your intentions to bring about certain outcomes as *ends,* for no further purpose. And this evidence that you will so act constitutes your *telic* reasons for action. This evidence is given to you by a certain history of action—namely, a history of both you and others acting to bring about certain outcomes as ends (e.g., your and others' well-being and autonomy). The "spring of action," then, on this view, is your instinctive disposition to induce from prior experience. Of course, this is just what we should expect when reducing the practical to the theoretical.

Once you have settled upon which end you are going to bring about, you initiate means-end practical reasoning—that is, the project of selecting an appropriate means to your end. For our radical cognitivist, means-end practical reasoning is just the project of inferring the most likely *causes,* given your evidence, of your bringing about those outcomes that you now predict you will attempt to bring about as ends—namely, your *acting* in certain ways that would, by your lights, help bring about said outcomes. The beliefs about what you will do, which this reasoning issues in, constitute your *instrumental* intentions—that is, your intentions to perform certain actions as *means* to bringing about your ends. This, in summary form, is how I propose that we can make sense of practical reason as a species of theoretical reason.

Of course, this sketch of the picture at hand needs much further elaboration and development. But it should give the reader a sense of the mechanics of the proposed reduction. When you are engaging in practical deliberation about what to do, what you are really doing is just trying to figure out what you will do, attending only to evidence about your future actions. Consequently, your intentions to act are revealed to really be certain beliefs about what you are going to do, and your reasons for action are unmasked as a kind of evidence concerning your future actions. And what it is to be engaged in means-end practical reasoning, it turns out, is really to be inferring the likely causes of your predicted future actions or the likely causes of the outcomes that you predict you will attempt to bring about, in light of your evidence.

Now that we have the basic picture under our belts, I want to bracket further consideration of the radical cognitivist's theory of practical reasons as evidence and instead focus our attention on her theory of intention and means-end practical reasoning. The bulk of the rest of this paper is devoted to showing how the radical cognitivist can construct a theory of intention and means-end practical reasoning, one that accommodates all the desiderata of any adequate theory of such phenomena but appeals only to the sparse resources—cognitive states, episodes of theoretical reasoning, and the epistemic norms governing them—to which she has limited herself.

However, before moving on, I first want to briefly draw attention to one point that is key to understanding the central significance of the radical cognitivist's theory of intention to her whole project. Radical cognitivism about practical reason does *not* entail that *any* evidence concerning what you will do counts as a reason for so acting—for example, that your evidence, gleaned from hard experience, that you will offend your host at their party constitutes a reason for you to offend them. No—that would be an absurd view. Rather, as was indicated above, only your evidence that concerns what outcomes you will act to bring about as an *end* (as a result, in the right kind of way, of this very evidence) is apt to constitute your telic reasons for action. And only your evidence that you will perform some action as a *means* to one of your ends (as a result, in the right kind of way, of this very evidence) is apt to constitute your instrumental reasons for action. This theory of practical reasons as evidence can, I claim, recover common sense about what we have reason to do—enough to be sensibly endorsed.[11]

But why should only this evidence count as your reasons for action? What *explains* this? Why should *any old* evidence concerning your future actions not count, on the radical cognitivist's reduction, as a reason for so acting? For example, why does your evidence that you will *F* not count, for the radical cognitivist, as a reason for you to *F*? After all, *if* the radical cognitivist held that your intention to *F* is just your belief that you will *F*, *then* she would be committed to the view that *any* evidence that you will *F* counts among your reasons to *F*. How so? Well, your reasons for action, by their nature, are just those considerations that count in favor of your forming an intention to act, and the considerations that count in favor of your forming the belief that you will *F* are, for the radical cognitivist, all and only your evidence that you will *F*. Consequently, if the radical cognitivist held this simple theory of intention, then she would be committed to any evidence that you will *F* as constituting a reason for you to *F*.

However, there is no reason to saddle the radical cognitivist with this particular theory of intention. First, this theory of intention would leave no room, in the radical cognitivist's picture, for the commonsense distinction between intending to *F* and merely foreseeing that you will *F*: your foresight that you will

[11] As I indicated before, here I focus just on the radical cognitivist's theory of intention and means-end reasoning. I return to the radical cognitivist's theory of practical norms as epistemic norms, and reasons for action as evidence, in a separate paper.

F is also just your belief that you will *F*, formed in light of evidence alone. Second, it is widely held among theorists of intention that the content of your intention to *F* is *not* simply the proposition "I will *F*."[12] Rather, it is broadly recognized that your intention to *F* has a more complex proposition as its content—namely, (something like) the proposition, "I will intentionally *F* as a result of this very mental state causing me in the right kind of way to intentionally *F*."[13] The radical cognitivist, I propose, should follow the orthodoxy in attributing to (the beliefs that constitute her) intentions this more complex content.

This move has a couple of important upshots. First, as I aim to show in this paper, it will allow the radical cognitivist to explain, purely through appeal to the sparse resources to which she has limited herself, how mere beliefs can occupy the functional role of intention: it is in virtue of their special content that the beliefs that constitute, on her view, intentions occupy a different functional role to the beliefs that constitute mere foresight. Second, it will straightforwardly allow the radical cognitivist to sidestep commitment to the (absurd) view that any evidence that you will *F* counts as one of your reasons to *F*. Instead, she will be committed to the view that the evidence that constitutes your reasons to *F* is just that evidence that counts in favor of your forming the belief that *you will intentionally F as a result of this very mental state causing you in the right kind of way to intentionally F*. And this evidence, I claim, is just your evidence that you will intentionally *F* as an end as a result, in the right kind of way, of (your awareness of) this very evidence, and your evidence that you will intentionally *F* as a means to one of your ends as a result, in the right kind of way, of (your awareness of) this very evidence. (I develop and defend this theory of practical reasons elsewhere.) In this way, then, simply by endorsing the standard view concerning the propositional content of intentions, the radical cognitivist can generate a psychologically adequate theory of intention and means-end practical reasoning and a normatively adequate theory of practical reason. And this, then, is why the radical cognitivist's theory of intention is of such central importance to her overall project: without it, she can neither explain how mere beliefs could occupy the functional role of intention nor explain why only that evidence concerning what you will do as an end or as a means to an end (as a result, in the right kind of way, of this very evidence)—and not any old evidence about your future actions—is apt to constitute your reasons for action. In other words, it is a structurally critical cornerstone of the radical cognitivist's theory of practical reason as a whole.

[12] Setiya, "Reasons without Rationalism"; Velleman, "Practical Reflection" and *Practical Reflection*. There are, of course, dissenters: Marušić and Schwenkler, for example, hold that your intention to *F* is just your belief that you will *F* ("Intending Is Believing").

[13] Harman, "Practical Reasoning."

2. Intention as Prediction, Planning as Inference

The contemporary orthodoxy in philosophical psychology on the nature of intention and means-end practical reasoning is best represented by the work of Michael Bratman.[14] For Bratman, intentions are commitments to action. In the same way that beliefs are theoretical commitments to the truth of a proposition, intentions are practical commitments to taking some course of action. They are the constituent elements of partial plans that get filled in (with further intentions and beliefs) as events unfold. Furthermore, for Bratman, they are also *sui generis* mental states, on a par with belief and desire and irreducible to them.

Nevertheless, the standard view still has it that we can characterize intentions and their status as commitments to action in terms of their having the following core features: intentions (1) are conduct controlling and (2) drive means-end practical reasoning.[15] What it is for a mental state to be conduct controlling is for it to be disposed to prompt you, at the appropriate time by your lights, to act in the way that it represents you as acting. And what it is for an intention to drive means-end practical reasoning, or planning, is for it to be such that it exerts rational pressure on you, at the right time by your lights, to plan out how you will act in the intended way.[16] I will follow the standard view here in assuming that it is a condition of adequacy on any theory of intention that it can accommodate these two features.

My principal goal in the rest of this paper is to show how the radical cognitivist can generate an adequate theory of intention—one that explains how mere belief can be both conduct controlling and drive means-end reasoning, and thus apt to occupy the functional role of intention purely through appeal to cognitive states, episodes of theoretical reasoning, and the epistemic norms governing them—simply by endorsing the standard view in the philosophical literature concerning the propositional content of intentions. As I said before, it is broadly agreed that the intention to *F* does *not* simply have the proposition "I will *F*" as its content.[17] Rather, it is generally thought to have as its content (something like) the proposition "I will intentionally *F* as a result of this very mental state causing me in the right kind of way to intentionally *F*."[18] The radical cognitivist, I claim, can generate an adequate theory of intention and means-

[14] Bratman, *Intentions, Plans, and Practical Reason.*

[15] Bratman, *Intentions, Plans, and Practical Reason*; Holton, *Willing, Wanting, Waiting.*

[16] And this *rational* pressure to begin planning will, in a sufficiently rational agent, constitute *motivational* pressure to begin planning.

[17] Setiya, "Reasons without Rationalism"; Velleman, "Practical Reflection" and *Practical Reflection.*

[18] Harman, "Practical Reasoning."

end reasoning just by attributing (the beliefs that constitute her) intentions this more complex content. She need not—as the moderate cognitivist about practical reason does—appeal to *sui generis* practical states or norms.

2.1. The Radical Cognitivist's First Pass at a Theory of Intention and Planning

Let us begin by considering the "primordial," or "first pass," formulation of the radical cognitivist's theory of intention and means-end practical reasoning. Consider the following instance of practical reasoning: it is midday and you form the intention to eat a burrito for lunch. You believe that you can eat a burrito if you walk to the food truck, purchase a burrito, and bite into it. In light of all this, you then form the intention to walk to the food truck, purchase a burrito, and bite into it. This intention then moves you to do just that. How do radical cognitivists propose that we make sense of this practical episode as a wholly theoretical enterprise involving only cognitive mental states governed solely by epistemic norms?

Radical cognitivists about practical reason—just like certain proponents of moderate cognitivism about practical reason—conceive of means-end reasoning as an instance of theoretical inference.[19] First, your intentions to act are identified with predictions about what you are going to do. And second, means-end reasoning is held to be a sequence of theoretical inferences concerning the likely causes of your predicted future actions. Thus, means-end practical deliberation commences after you have made a prediction about what you are going to do (say, eat a burrito) in light of your evidence. You know that the best explanation—or most likely cause—of your acting in that way is that you act in certain other ways (namely, that you walk to the food truck, purchase a burrito, bite into it, etc.). You consequently infer that you will act in those ways. You know that the best explanation of your acting in these various predicted ways is, ultimately, that your muscles contract in certain sequences. (This is where your reasoning transitions from the conscious, personal level to the unconscious, subpersonal level.) In light of this, you form an unconscious and subpersonal prediction about how your muscles are just about to contract. This prediction then causes your muscles to contract in that sequence that it predicts they will contract, and this in turn causes you to act in the way you predicted that you would. So, the practical deliberation at work in moving from intention (beliefs about your future actions) to motor command (predictions about muscle contractions) turns out to be an inference about the likely causes (muscle contractions) of predicted future states of the world (your act of eating a burrito).[20]

[19] Setiya, "Reasons without Rationalism."

[20] One worry the skeptical reader may have had concerning this theory of means-end reasoning is how it accounts for Buridan cases, in which two actions are equally good means to the end in question. Translated to the case of radical cognitivism about practical reason,

In essence, advocates of radical cognitivism about practical reason propose that, during such practical deliberation, you treat your intended end state as "observed" and then infer backward the most likely cause of your ending up in that state. Thus, for such radical cognitivists, means-end practical reasoning is just a species of theoretical reasoning—in particular, a sequence of inferences to the best explanation. In the same way that perception is the endeavor to explain your sensory input through inferring the distal causes of that input—this being the dominant understanding of perception in cognitive science—means-end practical reasoning is the project of explaining the occurrence of your predicted future states of your person through inferring the most likely cause of them (ultimately, contractions of your muscles).[21] Means-end reasoning is a kind of backward-moving causal reasoning that terminates in cognitive states—motor predictions about how your muscles are just about to contract—that causally suffice for overt behavior.

This initial formulation of the radical cognitivist's view presents us with an explanation of how mere beliefs about what you are going to do can, on the radical cognitivist's reduction, occupy the functional role of an intention. The common-ground view in the philosophical literature is that your intentions are apt to drive planning since, in tandem with certain background beliefs, they exert rational pressure on you to start planning how you are going to act in the way they represent you as acting.[22] For example, your intention to *F*, taken together with your background belief that you will *F* only if you start planning how to *F* now, exerts rational pressure on you to immediately start planning how you are going to *F*. You face decisive rational pressure to either give up your intention to *F*, give up your background belief, or immediately start planning how to *F*. Granting that you have good reason to hold on to the former two attitudes, the psychic move that you are rationally required to make is to immediately start planning how to *F*.

this would be a situation in which my evidence indicates that I am equally likely to bring about my end *E* by means *A* or means *B*. Means-end reasoning, on the radical cognitivist's picture, would grind to a halt in such a situation—since the canons of epistemic rationality require that, if I have equally good reason to believe two inconsistent propositions, I abstain from judgment. Unfortunately, I lack the space to properly discuss this objection here, and return to it in a separate paper that focuses on the radical cognitivist's theory of practical norms (Ratoff, "Practical Reason as Theoretical Reason"). However, one option available to the radical cognitivist here is simply to deny the setup of the problem: it is in fact *never* the case that there are two means to our end that our evidence indicates we are equally likely to perform. There is in fact always some asymmetry, by our lights, between action *A* and *B* that renders us more likely to perform one or other of the actions.

[21] Friston, "A Theory of Cortical Responses."

[22] Bratman, *Intentions, Plans, and Practical Reason*.

The radical cognitivist can seek to reproduce this account: your belief that you will *F* is apt to constitute, for the radical cognitivist, your intention to *F* since, in tandem with certain background beliefs, it exerts rational pressure on you to start planning how you will *F*. After all, just like your intention to *F*, your belief that you will *F*, together with your belief that you will *F* only if you now start planning how to *F*, will exert rational pressure on you to immediately start planning. You will face decisive *epistemic* rational pressure to either give up your belief that you will *F*, give up your belief that you will *F* only if you now start planning how to *F*, or immediately start planning how you are going to *F*. Granting that you have good reason to hold on to your first two beliefs, the psychic move that minimally mutilates your web of belief, and that you are therefore required by epistemic rationality to make, is to immediately start planning how you will *F*. After all, given your web of background beliefs, if you do not now start planning how you are going to *F*, how can you rationally continue believing that you will *F*? Of course, for the radical cognitivist, the project of planning how you will *F* is just the enterprise of inferring the most likely causes of your *F*-ing. This enterprise—if completed—will bottom out in motor predictions that will causally suffice for you to *F* in the way detailed by your plan. In this way, then, the radical cognitivist proposes to explain how mere beliefs about what you are going to do can be both conduct controlling and plan driving, and consequently apt to occupy the functional role of intention on the radical cognitivist's psychology, purely though appeal to cognitive states, episodes of theoretical reasoning, and the epistemic norms governing them.[23]

One immediate problem with the theory so far: the radical cognitivist holds that your predictions about how your muscles are just about to contract causally suffice for their predicted muscle contractions to occur. But this seems to be obviously false: clearly, you can predict that you are just about to contract your muscles without this prediction then causally sufficing for the contraction of your muscles. For example, suppose that the evil scientist is now directly stimulating the muscles in your arm such that they spasm and contract, causing you to move your arms around. The scientist looms over you, ready to stimulate your muscles again. You consequently infer that your muscles are just about to contract. The "primordial" radical cognitivist is committed to this prediction causally sufficing for your muscles to contract. So the radical cognitivist's picture seems to (implausibly) predict that you will intentionally move your muscles here, rather than merely foreseeing that they will move as a consequence of the scientist's stimulation.

[23] As I indicated above, this is just a "first pass" at the radical cognitivist's theory of intention and means-end reasoning and not the final product to be defended here. I take this formulation to be inadequate for reasons that will become clear.

The best response available to the radical cognitivist here, I think, is to appeal to reflective predictions. Harman, Setiya, and Velleman hold that your predictions about what you are going to do that constitute intentions, not mere foresight, are those that represent themselves as being the cause of the future actions that they represent.[24] You count as intending to *F* only when your prediction that you will *F* also represents itself as being the cause of your actually doing *F*. Let us follow Velleman in calling such beliefs "reflective predictions."

I propose that our radical cognitivist can immunize herself against the above counterexample by joining the above cognitivists about intention in holding that your motor commands are reflective predictions about what you are going to do. How does this pertain to the problem at hand? Well, she can hold that only your *reflective* motor predictions about how your muscles are just about to contract are causally sufficient for the occurrence of their predicted muscle contractions. Your prediction that your muscles are just about to contract in the *M* way is *not* causally sufficient for your muscles then contracting in that *M* way. No—only your prediction that *your muscles are just about to contract in the M way because of this very prediction* will causally suffice for your muscles to contract in the *M* way. And you arrive at these reflective motor predictions through the same backward-moving (likely unconscious) causal reasoning (that is constitutive of means-end practical reasoning on the radical cognitivist's account) through which you were theorized to arrive at a nonreflective motor prediction. In particular, you infer your reflective prediction that your muscles are just about to contract in the *M* way because of this very prediction from your (nonreflective) prediction that your muscles are just about to contract in the *M* way together with your belief that the best explanation—or most likely cause—of your muscles contracting in the *M* way, given your evidence, is that they will be caused to contract in the *M* way by your prediction that they are just about to contract in the *M* way. In this way, then, your reflective motor predictions, which causally suffice for their represented muscle contractions, are inferred to be the best explanation, given your evidence, of your predicted muscle contractions.

So, the radical cognitivist should adjust her theory of motor commands and hold that only reflective motor predictions are causally sufficient for their predicted muscle contractions. This allows her to explain why your prediction that your muscles are just about to contract in the evil-scientist case does not count as a motor command or intention: here you do *not* believe that this prediction is the (most likely) cause of your future muscle contractions. Rather, you believe that those muscle contractions will be the result of the scientist's

[24] Harman, "Practical Reasoning"; Setiya, "Reasons without Rationalism"; Velleman, "Practical Reflection" and *Practical Reflection*.

stimulations. Consequently, the rational pressure you would otherwise face to infer a reflective motor prediction from your nonreflective motor prediction is absent. Hence, you do not infer one. And your nonreflective motor prediction is not causally sufficient for its predicted muscle contractions. In this way, then, the radical cognitivist can now maintain (correctly) that in the evil-scientist case your arms will twitch not intentionally but rather as a result of the scientist's stimulation.

2.2. How Can Mere Belief Occupy the Functional Role of Intention?

A second problem facing the primordial radical cognitivist's theory of intention and means-end practical reasoning is that it fails to adequately explain how mere beliefs can occupy the functional role of intention. Recall that the characteristic features of intention are that it is conduct controlling and drives means-end practical reasoning. Thus, the radical cognitivist will have successfully shown how mere belief can occupy the functional role of intention just when she has shown how mere beliefs about what you are going to do can be both conduct controlling and also such that they exert rational pressure on you to plan out how exactly you will act as they indicate you will act.

But so far the radical cognitivist has failed to posit any (intrinsic) difference between those beliefs about what you will do that constitute, in her picture, your intentions to act and those beliefs that rather constitute your mere foresight about what you are going to do. According to the primordial radical cognitivist, your intention to *F* is simply a belief that you will *F*. But your mere foresight that you will *F* must also just be a belief that you will *F*. Given this, nothing could explain, for the primordial radical cognitivist, why your intention to *F* is both conduct controlling and such that it exerts rational pressure on you to plan out how you will *F* but your mere foresight that you will *F* possesses neither of these powers. What differentiates this former belief, which is supposed to constitute an intention, from the latter one, which does not? Where is the asymmetry?

Now, the moderate cognitivist about practical reason can explain the difference between those beliefs that constitute your intentions and those that rather constitute your mere foresight through appeal to a distinctively practical genealogy. For the sake of vivid illustration, consider the following concrete case. You are attending a fancy party. You aim at being entertaining. You therefore decide to tell a risqué joke, knowing that it will bring the house down. However, you also know that your host is a priggish prude who will certainly take offense at your joke. All things being equal, you would prefer not to offend your host. But you really want to amuse everyone else. Consequently, after you have weighed things up again you decide to go ahead and tell the joke anyway.

Intuitively, you here count as intending to tell a joke and merely foreseeing that you will offend your host.

Harman, Setiya, and Marušić and Schwenkler—moderate cognitivists all—theorize that the distinction between the beliefs that constitute your intentions to act and the beliefs that constitute your foresight about what you are going to do is that the former, but not the latter, are held in light of and made rational by practical reasoning—where practical reasoning is held to be *sui generis* and irreducible to any kind of theoretical reasoning.[25] As Marušić and Schwenkler put it: "intentions are beliefs—beliefs that are held in light of, and made rational by, practical reasoning."[26] So beliefs about what you are going to do count as foresight, on this account, when they are held purely in light of evidence, whereas such beliefs count instead as intentions when they are held in light of, and made rational by, practical reasoning—that is, the process of weighing the considerations for and against some course of action in light of your *sui generis* (believed) reasons for action or your desires and means-end beliefs. Since the latter set of beliefs is the product of practical reasoning, they are apt, these cognitivists claim, to be identified with intentions. So, your belief that you will tell a joke constitutes an intention to do so because it was formed in light of and rationalized by practical reasoning: you concluded that you would tell a joke in light of your believed reasons to amuse your audience and your belief that you could amuse them by telling them that joke. But your belief that you will offend your host counts as mere foresight, on this account, since it was formed in light of evidence, not practical reasoning: you inferred that you would offend your host as a causal upshot of your predicted act of telling the risqué joke. In this way, then, through appeal to a certain practical genealogy, the cognitivist can accommodate the commonsense distinction between intention and foresight.

However, this genealogical theory of the distinction between intention and foresight will not be available to the radical cognitivist. After all, she denies the reality of any *sui generis* practical states, reasoning processes, or norms, and holds that all practical reasoning is just an instance of theoretical reasoning concerning what you are going to do. Consequently, for the radical cognitivist, your beliefs about your future actions that constitute intentions, no less than those that constitute foresight, are formed purely in light of evidence. She therefore cannot hold that the distinction between intention and mere foresight is to be drawn genealogically, with the former, but not the latter, being the product of and rationalized by *sui generis* practical reasoning or norms.

[25] Harman, "Practical Reasoning"; Setyia, "Practical Knowledge"; Marušić and Schwenkler, "Intending Is Believing."

[26] Marušić and Schwenkler, "Intending Is Believing."

How might the radical cognitivist go about explaining how certain beliefs about what you are going to do, but not others, can occupy the functional role of intention? In other words: How can she accommodate the distinction between intention and foresight? One natural thought is that the answer is already on the table: your intentions, the radical cognitivist can hold, are just your *reflective* beliefs about what you are going to do. So, perhaps the radical cognitivist should hold that the difference between intention and mere foresight resides in whether or not the prediction in question represents itself as the cause of its predicted future action. In this picture, your reflective predictions are just your intentions to act, with your nonreflective predictions about your future actions instead counting as mere foresight.

This adjustment to her theory is *not ad hoc*. There are compelling reasons for any cognitivist about intention to identify intentions to act with reflective predictions. After all, it is common ground between all theorists of intention that intentions are formed to ensure that we act in the intended way once the time comes.[27] Hence, everyone will agree that when we intend to act we believe that we will act in the intended way, if indeed we do so act, as a causal consequence of our intention to act in that way. After all, had we believed that we would act in that way as a causal upshot of something *other* than our intention, then we would not have judged it necessary to form an intention in the first place. As Velleman puts it,

> the content of an agent's intention of doing something cannot be merely that he's going to do it, because of some impetus or other; it must be that he is going to do it partly because of this very intention. If the agent could intend to do something, without intending to do it partly because of so intending, then he could intend to do the thing unintentionally—which he can't.[28]

And Setiya—another cognitivist—agrees:

> Intention is self-referential. When I intend to ϕ, the content of my attitude is that I am going to ϕ because of that very intention: intention represents itself as motivating action.... It is part of what one believes in deciding to do something that one's choice will be efficacious; without that belief, decision would make no sense.[29]

[27] Harman, "Practical Reasoning"; Bratman, *Intentions, Plans, and Practical Reason* "; Velleman, *Practical Reflection.*

[28] Velleman, *Practical Reflection.*

[29] Setiya, "Reasons without Rationalism."

In short, when you form an intention to *F*, your intention to *F* will represent itself as being the cause of your future *F*-ing. All parties to the debate should agree. Consequently, if your intentions are beliefs about what you will do, as the cognitivist insists, then they must be reflective beliefs: they must represent themselves as the cause of their predicted future actions. Given this, the radical cognitivist, too, should hold that your intentions to act are reflective beliefs about what you are going to do.

However, this way of drawing the distinction between intentions and mere foresight is not going to work. To see why, let us consider a popular counterexample in the literature to the thesis that your intentions are just your reflective predictions: Bratman's case of the pessimistic actor.[30] The pessimistic actor believes that he will stumble over his lines and that he will stumble over his lines as a result of this very belief. Perhaps he has a neurosis that he tends to focus too much on self-prediction and not enough on just saying his lines such that he believes this inappropriate focus will cause him to actually fluff his delivery of his lines. So, the pessimistic actor believes that he will stumble over his lines as a result of this very belief. But he does not intend to stumble over his lines. Quite the opposite! He intends to deliver them appropriately. Hence, there must be more to your intentions than mere reflective beliefs about what you are going to do.

The case of the pessimistic actor demonstrates how mere reflective beliefs about your future actions are not apt to occupy the functional role of an intention in the radical cognitivist's psychology. Your intention to *F* by its very nature necessarily exerts rational pressure on you to start—at the right time, by your lights—planning out how to *F*.[31] But your belief that you will *F* as a result of this very belief does not, even on the radical cognitivist's psychology, necessarily exert rational pressure on you to start planning out how you will *F*. As the case at hand illustrates, it is perfectly possible for you to believe that you will *F* as a result of this very belief but fail to face any rational pressure to begin planning out how you will *F*. After all, the pessimistic actor believes that he will stumble over his lines *without* him having to plan out how he will so stumble. He therefore faces no stark choice between giving up his reflective belief that he will stumble over his lines and starting to plan how he will do so. Hence, his reflective belief that he will stumble over his lines, unlike an intention to do so, does not exert any rational pressure on him to plan out how he will so act. In this way, then, we can see how mere reflective beliefs about what you are going to do are not apt, even on the radical cognitivist's psychology, to

[30] Bratman, "Cognitivism about Practical Reason."

[31] Bratman, *Intentions, Plans, and Practical Reason*.

occupy the functional role of intention: they neither drive planning nor count as conduct controlling.

In light of this problem, the radical cognitivist ought, I think, to further develop her theory of intention. As I indicated before, the radical cognitivist can generate an adequate theory of intention simply by attributing (the beliefs that constitute) her intentions the propositional content that the standard view in the philosophical literature assigns to intentions—namely, that the intention to *F* has the content "I will intentionally *F* as a result of this very mental state causing me in the right kind of way to intentionally *F*."[32] Given that this is my strategy, I now want to remind the reader of why this standard view of the propositional content of intention is broadly accepted.

2.3. *The Standard View of the Content of Intention*

It is common ground among many theorists of intention not just that intentions represent themselves as the causes of their predicted future actions but also that intentions, by their nature, represent their predicted future actions as being caused by themselves *in a certain kind of way*. What kind of way? Your intention to *F*, it is broadly agreed, represents itself as causing you to *F*, first, by exerting rational pressure on you to plan how you will *F* and, second, by this process of planning bottoming out in motor commands that cause you to *F* in the way that your plan detailed.[33] The standard theory of intentional action in the literature—the causal theory of intentional action—has it that you count as intentionally *F*-ing just when you *F* as a causal consequence of your project of planning out how to *F* having bottomed out in motor commands that cause you to *F* in the way detailed by your plan.[34] Hence, we can more concisely articulate this second constitutive feature of intention by substituting this theory of intentional action into the content of an intention—thus: your intention to *F* represents itself as causing you to intentionally *F* by exerting rational pressure on you to intentionally *F*.

Why think this? Why join proponents of the standard view in thinking that intentions have this representational content? Well, intentions, according to the standard view, have a certain constitutive "world-mind" direction of fit: they aim at making you change the world such that it "fits" the content of your intentions.[35] This content represents the "success condition" of the intention: the condition that must obtain for the intention to count as having satisfied its

[32] Harman, "Practical Reasoning."

[33] Harman, "Practical Reasoning."

[34] Harman, "Practical Reasoning"; Paul, "Deviant Formal Causation."

[35] Smith, *The Moral Problem*.

constitutive aim. Given this, we can infer the content of an intention simply by figuring out its success condition. And the success condition of your intention to *F*, it turns out, is your intentionally *F*-ing as a result of this very intention to *F* exerting rational pressure on you to do so. Hence, it follows that your intention to *F* must represent itself as causing you to intentionally *F* by exerting rational pressure on you to intentionally *F*.

The key question is this: Why think that the success condition of your intention to *F* is your intentionally *F*-ing as a result of this very intention to *F* exerting rational pressure on you to intentionally *F*? The answer: because common sense suggests that this is the condition under which your intention to *F* counts as having satisfied or achieved its constitutive aim. This should be evident after consideration of a few concrete cases. First, it is clearly *not* enough for your intention to *F* to count as having satisfied its constitutive aim for it to have merely caused you, in one way or another, to *F*. No—we regard your intention to *F* as having fallen short of its aim if it prompts you to *F unintentionally*. Consider the following example: you intend to kill someone by shooting him. However, when you shoot, you miss by a mile. But your shot causes a herd of wild pigs to stampede such that they trample your intended victim to death. Here your intention to kill the man is indeed the cause of your killing him. But it does not cause you to kill him "in the right kind of way" for you to count as having *intentionally* killed him.[36] Rather, for your intention to *F* to count as causing you to intentionally *F*, it must have caused you to *F* by, first, causing you to plan out how to *F* ("I will kill him by shooting him dead") and this process of planning then causing you to *F* in the way that your plan details.[37] Furthermore, and most pertinently for us, intuition suggests that your intention to kill him did not satisfy its constitutive aim here. This is good evidence that, in general, your intention to *F* only counts as having satisfied its constitutive aim if it causes you to *intentionally F*. Hence, we should think that your intention to *F* must represent itself not just as the cause of your *F*-ing but also as causing you to intentionally *F*.

Second, it is not enough for your intention to *F* to count as having satisfied its constitutive aim for it to have caused you to intentionally *F*—that is, to *F* as a result of planning how to *F* and in the way detailed by your plan—in *any old way*. Rather, it must have caused you to intentionally *F in the right kind of way*—namely, by exerting *rational pressure* on you to start planning how to *F* and this process of planning then bottoming out in a way that causes you to *F* in the way detailed by your plan. We regard your intention to *F* as having failed

[36] Davidson, "Freedom to Act."

[37] Harman, "Practical Reasoning"; Paul, "Deviant Formal Causation."

to satisfy its constitutive aim if it causes you to intentionally *F* in some way *other* than through it having exerted rational pressure on you to plan out how to *F*. Take the following case: you are extremely busy at work over the Christmas period and you are unsure whether you ought to fly home for Christmas. After some reflection, you decide that you will fly home after all. You consequently form an intention to book flights home for Christmas. However, given all the cognitive pressures you are under, you soon forget all about your intention to do this. Nevertheless, your earlier awareness of your intention causes a chain of festive thoughts that ends up causing you to weigh up the reasons for and against flying home for Christmas. In light of this practical deliberation, you then form a (second) intention to book flights home, one that prompts you in the normal way to actually do so—that is, by exerting rational pressure on you to plan how to bring it about that you book said tickets and that process eventually causing you to book the tickets in the way detailed by your plan. Here your initial intention to book flights home for Christmas did indeed cause you—via a deviant causal chain—to intentionally book flights home. But, intuitively, this intention did not satisfy its constitutive aim. Only your second, later intention to book flights home—the one that caused you to do so by exerting *rational pressure* on you to plan how to do so, and so on—seems to have satisfied its constitutive aim. This is good evidence that, in general, your intention to *F* only counts as having satisfied its constitutive aim if it causes you to intentionally *F in the right kind of way*—that is, by exerting *rational pressure* on you to plan out how to *F* and this process of planning then causing you to *F* in the way detailed by your plan. In other words, granting the truth of the standard causal theory of intentional action, your intention to *F* only counts as having achieved its constitutive aim if it causes you to intentionally *F* by exerting *rational pressure* on you to intentionally *F*. Hence, we should think that your intention to *F* must represent itself not just as the cause of your intentionally *F*-ing but as causing you to intentionally *F* by exerting *rational pressure* on you to intentionally *F*.

2.4. The Radical Cognitivist's Theory of Intention

We have now seen why it should be agreed that your intention to *F* by its nature represents itself as causing you to intentionally *F* through exerting rational pressure on you to intentionally *F*. This is simply (a precisification of) the standard view in the literature concerning the propositional content of intentions—namely, that my intention to *F* has the content "I will intentionally *F* as a result of this very mental state causing me in the right kind of way to intentionally *F*."[38] This points the way for the radical cognitivist: your

[38] Harman, "Practical Reasoning."

intention to *F* cannot just be your belief that you will *F*. No—it must be your belief that you will intentionally *F* as a result of this very mental state exerting rational pressure on you to intentionally *F*. Let us call this belief your "rationally reflective prediction that you will intentionally *F*." We say that it is reflective because it represents itself as the cause of your intentionally *F*-ing. So, I say that it is rationally reflective because it represents itself not just as the cause of your intentionally *F*-ing but also as the cause of your intentionally *F*-ing *in a certain kind of way*—namely, through exerting *rational pressure* on you to intentionally *F*. This, I claim, is the correct formulation of the radical cognitivist's theory of intention: your intention to *F* is just your belief that you will intentionally *F* as a result of this very belief exerting rational pressure on you to intentionally *F*. More concisely: your intention to *F* is just your rationally reflective prediction that you will intentionally *F*.

> *Intention:* S intends to $F =_{\mathrm{df}} S$ believes that S will intentionally F as a result of this very belief exerting rational pressure on S to intentionally F.

One obvious problem: this formulation entails that motor predictions concerning how your muscles are just about to contract—the radical cognitivist's candidate for motor commands—do not count as intentions since they are not rationally reflective. You do not believe that you need to plan out how you will perform the intended motor contractions in question. Rather, such motor commands simply causally suffice for the occurrence of their represented muscle contractions. However, this problem is easily solved through a small tweak to our theory: intentions—other than motor commands—are all rationally reflective predictions. And motor commands are just your *reflective* predictions about how your muscles are just about to contract—that is, your predictions that your muscles are just about to contract as a result of these very predictions.

A second problem: Does this radically cognitivist theory of intention not presuppose the notion of an intentional action and thus of a plan? In order to (noncircularly) theorize intentions in terms of the notion of an intentional action or a plan, we must already have a prior understanding of intentional actions and plans that makes no reference to intention. Of course! But the radical cognitivist can analyze intentional action and planning in wholly cognitivist terms without reference to the notion of an intention. First, she can follow the standard view on the nature of intentional action—the causal theory—in holding that you count as intentionally *F*-ing just when you *F* as a causal consequence of your planning out how to *F* and in the way detailed by your plan to *F*.[39] And, as we saw before, for the radical cognitivist, what it is for you to

[39] Paul, "Deviant Formal Causation."

be planning how you will *F* is just for you to be inferring the best explanation of your predicted act of *F*-ing where this reasoning would conclude—if completed—in a reflective motor prediction that suffices for action. The radical cognitivist faces no circularity here. Thus, the radical cognitivist can hold that your intention to *F* represents itself as causing you to *F* by prompting you to begin inferring the causes of your *F*-ing, a process that bottoms out—if completed—in action. This is what it is, for the radical cognitivist, for your intention to *F* to represent itself as causing you to intentionally *F*.

Can this formulation of the radical cognitivist's theory of intention explain how a mere belief can occupy the functional role of intention—that is, be both conduct controlling and plan driving—purely through appeal to cognitive states, episodes of theoretical reasoning, and epistemic norms? Can it correctly class instances of genuine intention, by the lights of common sense, as intention and the cases of mere foresight as foresight? I think so. This "rationally reflective" content suffices to render a belief, in the radical cognitivist's austere psychology, both conduct controlling and plan driving.

How precisely does this work? Take the earlier case: you are attending a fancy party thrown by a host who is a well-known prude. You decide to tell a risqué joke, aiming to entertain your audience, while knowing that it will offend your host. Intuitively, you count as intending to tell the joke but merely foreseeing that you will offend the host. According to the radical cognitivist, your intention to tell the joke is just your rationally reflective prediction that you will intentionally tell the joke. This state is apt to constitute your intention here because it exerts decisive rational pressure on you, in concert with the right background beliefs, to start planning out how you will tell the joke, and thereby counts as conduct controlling and plan driving.

Let us break down how this is supposed to go. You believe that you will tell the joke, that you will tell the joke as a result of this very belief, and that this belief will cause you in the right kind of way to *intentionally* tell the joke—namely, that it will first cause you in the right kind of way to starting planning out how you will tell the joke by exerting rational pressure on you in tandem with background beliefs to begin planning, and, second, that this planning will then cause you to tell the joke in the way it details. This is the content of your rationally reflective prediction spelled out. Now, how does this rationally reflective prediction prompt you, at the appropriate time by your lights, to start planning out how you will tell the joke? Well, suppose that you believe that you will only intentionally tell the joke, as an upshot of your prediction that you will intentionally tell the joke, causing you in the right kind of way to do so, if you start planning right now how you are going to tell it. Granting this, given the norms of theoretical rationality, you must (rationally) either give up your rationally reflective prediction that you will intentionally tell the joke or start planning out how you will tell it.[40] Suppose again that you have more reason to believe that you will intentionally tell the joke (as a causal upshot, in the right kind of way, of this very belief) than you have to believe that you will *not* start now planning out how you will tell it. Now the psychic move that does the least epistemic violence to your web of beliefs—and that you are therefore required by epistemic rationality to make—is to begin inferring the causes of your telling the joke (that is, to start planning). Consequently, you will—insofar as you are rational—start planning out how you will tell the joke. This process of planning will—if completed—bottom out in reflective motor predictions that will cause you to tell the joke in the way detailed by your plan. Hence, your rationally reflective prediction that you will intentionally tell the joke counts as both plan driving and conduct controlling. This is why this rationally reflective prediction is apt to constitute your intention to tell the joke.

The *reflective* character of your prediction is playing an important role here: if you believed that you would tell a joke without this action being caused by your prediction that you would do so, then your prediction that you would tell this joke would not exert any rational pressure on you to start planning out how you will tell it. After all, you would believe that you would tell the joke as a result of some other impulse, without this very belief prompting you to plan out how. You would therefore face no stark choice between giving up your belief that you would tell the joke or starting to plan out how to tell it. You could rationally hold on to your belief that you will tell the joke, yet fail to start planning out how. Hence, this belief exerts no rational pressure on you to start planning. In this way, then, we can see how the fact that your prediction is reflective is essential to it being apt to occupy the functional role of an intention.

So too is the fact that you predict that you will *intentionally* tell the joke. After all, if you believed (somehow) that you would *unintentionally* tell the joke as a result of your belief that you would tell that joke—like in the pessimistic actor case—then your belief would not exert any rational pressure on you to start planning out how to tell that joke. How so? Well, just like the pessimistic actor, you would believe that your belief that you will tell the joke will causally suffice alone for you to actually tell the joke without your having to plan out how to go about telling it. You would therefore face no stark choice between

[40] Or, you could give up your background belief that you will only intentionally tell the joke, as a upshot of your prediction that you will intentionally tell the joke causing you in the right kind of way to do so, if you start planning right now how you are going to tell it. Of course, if you have this background belief, then you are likely to be warranted in holding it, so it would likely be theoretically irrational of you to revise this belief. (I am omitting this caveat henceforth for ease of exposition.)

giving up your belief that you will intentionally tell the joke and starting to plan out how you will tell it. Hence, this belief exerts no rational pressure on you to start planning. In this way, then, the fact that your prediction represents your predicted action as being *intentional* is also essential to that prediction being apt to constitute an intention.

Last, the fact that your prediction is *rationally* reflective in character is also pertinent. If you believed that you would intentionally tell the joke as a result of your belief that you will intentionally tell the joke causing you to plan out how to do so in a *deviant* way—say, by prompting you to consider the reasons for telling a joke, and so on, like in the case of your booking flights home for Christmas—then said belief would not exert any rational pressure on you to begin planning. Why? Well, there would be no rational pressure to begin planning exerted by your reflective belief here since your predicted act of intentionally telling the joke is accounted for in a way—namely, the deviant way—that does not involve your reflective prediction that you will intentionally tell the joke causing you in the right kind of way to tell it—that is, through rationally pressuring you to plan out how you will tell it. You can rationally hold on to this merely reflective prediction that you will intentionally tell the joke while abstaining from planning out how you are going to tell it. Hence, this merely reflective prediction about what you will intentionally do exerts no rational pressure on you to start planning. In other words, the fact that your reflective prediction that you will tell the joke represents itself as causing you to tell the joke *in the right kind of way*—through exerting rational pressure on you to plan, and so on—is essential to this prediction being apt to constitute an intention. In short, your prediction being *rationally* reflective is necessary for it being such that it can exert rational pressure on you to begin planning and thus being apt to constitute an intention. And, in sum, a rationally reflective prediction that you will intentionally *F* is necessary and sufficient, on the radical cognitivist's psychology, for you to be in the kind of plan-driving and conduct-controlling state that is apt to constitute an intention to *F*.

Let us now turn to your mere foresight that you will offend your host. Can the radical cognitivist accommodate this? Yes—according to the radical cognitivist, this prediction counts as mere foresight because it is not a rationally reflective prediction concerning what you will do. Indeed, it is not even a *reflective* prediction: you do not believe that you will offend your host as a result of your belief that you will offend him. You believe that you will offend your host, even if you do not expect to offend him. Your rationally reflective belief that you will tell a joke will take care of that. Nor do you believe that you will offend your host as an upshot of planning out how you will offend him. On the contrary, you believe that you will offend your host as a causal upshot of some other action (telling the risqué joke) you are planning. Hence, your prediction that you will offend your host exerts no rational pressure on you to start planning how to do that: you face no stark choice between giving up your belief that you will offend your host or starting planning how you will do it. In this way, the radical cognitivist can correctly class this prediction as an instance of mere foresight, not intention.

We have now seen how the radical cognitivist can explain how mere belief can occupy the functional role of intention purely through appeal to the sparse resources to which she has limited herself—namely, by attributing (the beliefs that constitute) her intentions the same propositional content the standard view on the nature of intention does. This account also allows the radical cognitivist to accommodate the commonsense distinction between intention and mere foresight. I now want to draw attention to the fact that the radical cognitivist's theory of intention accommodates another key element of the common ground on the nature of intention—namely, the distinction between your instrumental intentions and telic intentions. Now, you *instrumentally* intend to *F* just when you intend to perform action *F* as a *means* to bringing about some end *E* that you already intend to bring about. And your intention to *F* is *telic* just when you intend to perform action *F* as an *end*, for no further purpose. The radical cognitivist can make sense of this distinction in her own terms. First, she can say that you *instrumentally* intend to *F* just when (1) you rationally reflectively believe that you will intentionally *F* and (2) this belief is warranted in light of your means-end belief that you can (help) bring it about that you *G* by *F*-ing and your rationally reflective belief that you will intentionally *G*. Second, she can say that you have a *telic* intention to *F* just when (1) you have a rationally reflective belief that you will intentionally *F* and (2) this belief does *not* constitute, for the radical cognitivist, an instrumental intention to *F*. It clearly follows from these definitions that for the radical cognitivist every intention is either instrumental or telic. In this way, then, the radical cognitivist can recover the mutually exclusive and exhaustive partition of intentions into their instrumental and telic varieties that is recognized by the common ground on the nature of intention.

3. Intention and the Orientation of Practical Reason

I want to conclude by considering one last problem the radical cognitivist's theory of intention might be thought to face, which concerns the forward-looking orientation of practical reason. It is part of the common ground that practical reasoning commences with you attending to the outcomes you could bring about. Indeed, practical reasoning seems by its nature to involve *only* consideration of the likely upshots or the intrinsic features of the actions available to you. It is essentially forward looking in nature. This contrasts with theoretical reasoning, which is often backward looking in orientation: "Why am I so confident that the sun will rise tomorrow? Because in my past experience it has risen every day." In sum, practical reason seems by its nature to involve only contemplation of future states of affairs—those that might be brought about by action—whereas theoretical reason is not restricted in this way: theoretical conclusions concerning the future can be arrived at after attention only to states of affairs that obtained in the past.

The radical cognitivist's conception of practical reason as a species of theoretical reason might therefore be thought to face difficulties accounting for the forward-looking orientation of practical reason. After all, if practical reason is just a branch of theoretical reason, and if theoretical reason can be backward looking in orientation, then why should practical reason be essentially forward looking in nature? What, for the radical cognitivist, could explain this? Restricting ourselves here just to her theory of intention, it looks like it is part of the common ground on the nature of intention that intentions to act are by their nature only held in light of forward-looking considerations concerning the intrinsic features or likely upshots of the intended action in question. You only intentionally act when you take that action to help bring about some outcome that you have taken as your end, for the sake of which you are performing that action. No rational agent ever intends to perform some action without taking it to promote their ends. This is part of the common ground on the nature of intention in philosophical psychology.

But beliefs, in contrast, can be held in light of backward-looking considerations. This remains as true for your rationally reflective beliefs about what you are going to intentionally do as it is for any of your other beliefs. This should lead us to doubt whether the radical cognitivist can accommodate the platitude that intentions are by their nature only held in light of forward-looking considerations concerning the intrinsic features or likely upshots of the intended action in question. By way of illustration, consider the following case concerning a seer's prophesy: the radical cognitivist is committed to holding that intentions are beliefs with a certain special content *P*. But surely, for any arbitrary content *P*, a reliable—by your lights—seer could inform you that *P* is the case. In that case, according to the radical cognitivist, you will—if rational—form an intention to act in light of this testimony. But that seems absurd: a seer's prophesy can at most warrant you to form mere foresight. To take the radical cognitivist's rationally reflective theory of intention that I am hawking here: suppose that a reliable seer, by your lights, informs you that you will intentionally kill your father as a result of your belief that you will intentionally murder him, which you will now form in light of this very prophesy, causing you to intentionally murder him in the right kind of way—namely, by rationally pressuring you to intentionally murder him. (In other words, the seer informs you that the content of a rationally reflective belief that you will intentionally murder your father is true). This prophesy of the seer, given your background belief that the seer is reliable, warrants you to form the rationally reflective prediction that you will intentionally murder your father. Suppose that, being rational, you now form this rationally reflective prediction. Radical cognitivism about practical reason now seems to imply that you have formed with warrant the intention to murder your father. But this seems absurd: in light of this testimony, you are at most warranted to form the mere foresight that you will murder your father, not an intention to do so. The radical cognitivist is failing to accommodate the platitude that intentions are essentially only held in light of forward-looking considerations and not backward-looking ones such as the seer's testimony.

However, I think that the radical cognitivist has the resources to accommodate common sense here—namely, that the seer's testimony that you will murder your father could not warrant you to form an intention to do just that, and, more generally, that you can rationally intend to perform some action only if you believe that so acting will help bring about one of your ends. How might she go about establishing this? Well, you have a lifetime of evidence that you only ever perform actions as ends or else as means to some further end.[41] Consequently, you cannot rationally believe that you will perform an action as anything other than as an end or else as a means. And this straightforwardly entails that you can rationally form the intention to kill your father in light of the seer's testimony only if you can rationally believe that you will perform this action as an end or as a means. But, as I shall argue, you cannot now, right after hearing the seer's testimony, rationally believe that you will kill your father as an end. Consequently, you cannot now rationally form a *telic* intention to kill your father as an end. And you cannot now rationally form an *instrumental* intention to kill your father as a means to some further end either, because such an intention must be formed in a certain kind of way, a way that does not obtain in the case of the seer's testimony. Since all intentions are either telic or instrumental, it follows that you cannot rationally form an intention to murder your father in light of the seer's testimony *tout court*.

Why can you not now rationally believe that you will kill your father as an end, for no further purpose? Well, killing your father is simply *not* the kind of

[41] The radical cognitivist can say that you perform action *E* as an end just when you perform *E* as a result of your telic intention to *E*, and that you perform action *M* as a means to some further end just when you perform *M* as a result of your instrumental intention to *M*.

thing that you (or anybody, for that matter) would ever seek to bring about as an end—and you know it. It is the kind of thing that could only ever be a *means* to some further end—revenge for some past grievous wrong, or to save the life of your child, for example. What outcomes do you have a history of acting to bring about as an end? Speaking for myself, my whole life has been at bottom a combination of looking out for myself and looking out for others. My ends—the outcomes I pursue for no further purpose—have ultimately been just my self-interest, the good of others, and what morality requires of me. This is what my life has unerringly been. So, I think, has everyone else's life.[42] Consequently, it just does not make any sense to you that you will murder your father as an end. Hence, you cannot rationally believe that you will so act. Thus, for the radical cognitivist, you cannot in light of the seer's testimony rationally form a telic intention to kill your father as an end.

And why can you not, right after hearing the seer's testimony, rationally form an *instrumental* intention to kill your father as a means to some further purpose? Well, instrumental intentions are by their nature formed in light of a telic intention to bring about some outcome and a means-end belief that you can (help) bring about that outcome by performing the action that is the object of your instrumental intention. The radical cognitivist, as we saw before, is able to accommodate this in the following way: you instrumentally intend to *F* just when (1) you rationally reflectively believe that you will intentionally *F* and (2) this belief is warranted in light of your means-end belief that you can (help) bring it about that you *G* by *F*-ing and your rationally reflective belief that you will intentionally *G*. Now, as these accounts make clear, you can instrumentally intend to *F* only if you *formed* that intention in light of a telic intention to bring about some outcome and your means-end belief that you can (help) bring about that outcome by *F*-ing. But your belief that you will murder your father was *not* formed in such a way. You did not infer that you would murder your father from your rationally reflective belief that you would intentionally bring about some outcome *O* and your means-end belief that you could (help) bring about *O* by murdering your father. No—you formed this belief in light of the seer's testimony. Hence, this belief could not, on the radical cognitivists' account, constitute an instrumental intention to murder your father. In this

[42] Some caveats: many may have pursued ends that cannot be conceived as prudent, pro-social, or moral ends—for example, epistemic ends of acquiring knowledge for its own sake or religious ends such as the worship of God. Very plausibly, many theists may take the worship of God to be an end that is performed for neither their own self-interest, the good of others, nor anything falling under the dominion of morality. And many (professional and nonprofessional) philosophers, scientists, and historians, etc., may pursue knowledge as an end and not for the sake of their prudence, etc.

way, then, we arrive at the conclusion that the seer's prophesy cannot, in the radical cognitivist's picture, warrant you to form the instrumental intention to murder your father.

Now, since all intentions are either telic or instrumental, it follows that you cannot rationally form an intention to murder your father in light of the seer's testimony. Rather, the seer's testimony warrants you only to form the mere foresight that you will kill your father: given your web of beliefs, you cannot rationally believe him when he tells you that you will kill your father as a result (in the right kind of way) of *this* very belief that you now form in light of his testimony. Instead, you can only rationally form the (nonreflective) belief that you will kill your father as a result of some other belief that you will form at some later date. This accords with our commonsense intuitions about the case of the seer's prophesy. And this result generalizes: your lifetime of evidence that you only ever perform actions as ends or else as means to some further ends ensures that you cannot rationally believe that you will perform an action as anything other than an end or else a means. And this straightforwardly implies that a rational agent can never intend some course of action without taking it to promote one of her ends. In this way, then, the radical cognitivist can accommodate the forward-looking orientation of practical reason in the domain of the theory of intention.

4. CONCLUSION

This completes my attempt to develop and defend the radical cognitivist's theory of intention and means-end practical reasoning. Intentions in this picture are rationally reflective predictions about what you are going to intentionally do that exert rational pressure on you to start planning. And means-end reasoning is a species of inference to the best explanation of your predicted actions that terminates—if completed—in action. Unlike other cognitivists about practical reason, the radical cognitivist reduces practical normativity to a variety of epistemic normativity, and therefore faces unique challenges in accounting for the basic desiderata on any adequate theory of intention and means-end practical reasoning. Here I showed how mere beliefs can occupy the functional role of intention, and how means-end practical reasoning can be a species of theoretical inference, purely through appeal to cognitive states, episodes of theoretical reasoning, and the epistemic norms governing them.

Trinity College Dublin
william.je.ratoff@gmail.com

REFERENCES

Adler, Jonathan E. *Belief's Own Ethics*. Cambridge, MA: MIT Press, 2002.

Anscombe, G. E. M. *Intention*. Cambridge, MA: Harvard University Press, 1957.

Audi, Robert. "Intending." *Journal of Philosophy* 70 , no. 13 (July 1973): 387–403.

Bratman, Michael. "Cognitivism about Practical Reason." *Ethics* 102, no. 1 (October 1991): 117–28.

———. *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.

Clark, Andy. "Beyond Desire? Agency, Choice, and the Predictive Mind." *Australasian Journal of Philosophy* 98, no. 1 (2019) 1–15.

———. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press, 2015.

———. "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." *Behavioral and Brain Sciences* 36, no. 3 (June 2013): 181–204.

Davidson, Donald. "Freedom to Act." In *Essays on Actions and Events,* 137–56. Oxford: Oxford University Press, 1973.

Davis, Wayne. "A Causal Theory of Intending." *American Philosophical Quarterly* 21, no. 1 (January 1984): 43–54.

Friston, Karl. "A Theory of Cortical Responses." *Philosophical Transactions of the Royal Society B* 360, no. 1456 (April 2005): 815–36.

Friston, Karl, Robert Adams, and Read Montague. "What Is Value—Accumulated Reward or Evidence?" *Frontiers in Neurorobotics* 6 (November 2012): 1–25.

Friston, Karl, James Kilner, and Lee Harrison. "A Free Energy Principle for the Brain." *Journal of Physiology–Paris* 100, nos. 1–3 (July–September 2006): 70–87.

Grice, H. P. "Intention and Uncertainty." *Proceedings of the British Academy* 57 (1971): 263–79.

Harman, Gilbert. "Practical Reasoning." *Review of Metaphysics* 29, no. 3 (March 1976): 431–63.

Hohwy, Jakob. *The Predictive Mind*. Oxford: Oxford University Press, 2013.

Holton, Richard. *Willing, Wanting, Waiting*. Oxford: Oxford University Press, 2009.

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

Marušić, Berislav, and John Schwenkler. "Intending Is Believing: A Defense of Strong Cognitivism." *Analytic Philosophy* 59, no. 3 (September 2018): 309–40.

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

Paul, Sarah K. "Deviant Formal Causation." *Journal of Ethics and Social Philosophy* 5, no. 3 (April 2011): 1–23.

———. "How We Know What We're Doing." *Philosophers' Imprint* 9, no. 11 (October 2009): 1–24.

———. "Intention, Belief, and Wishful Thinking: Setiya on 'Practical Knowledge.'" *Ethics* 119, no. 3 (April 2009): 546–57.

Ratoff, William. "Practical Reason as Theoretical Reason." Unpublished manuscript.

Ross, Jacob. "How to Be a Cognitivist about Practical Reason." In *Oxford Studies in Metaethics*, vol. 4, edited by Russ Shafer-Landau, 243–81. Oxford: Oxford University Press, 2009.

Setiya, Kieran. "Practical Knowledge." *Ethics* 118, no. 3 (April 2008): 388–409.

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

Shafer-Landau, Russ. *Moral Realism: A Defence*. Oxford: Oxford University Press, 2003.

Shah, Nishi. "A New Argument for Evidentialism." *Philosophical Quarterly* 56, no. 225 (October 2006): 481–98.

Smith, Michael. *The Moral Problem*. Oxford: Blackwell, 1994.

Velleman, J. David. "Practical Reflection." *The Philosophical Review* 94, no. 1 (January 1985): 33–61.

———. *Practical Reflection*. Chicago: University of Chicago Press, 1989.

Wallace, R. Jay. "Normativity, Commitment, and Instrumental Reason." *Philosophers' Imprint* 1, no. 3 (December 2001): 1–26.

Way, Jonathan. "Two Arguments for Evidentialism." *Philosophical Quarterly* 66, no. 265 (October 2016): 805–18.

Journal of Ethics and Social Philosophy
vol. 26, no. 1 · november 2023

https://doi.org/10.26556/jesp.v26i1.1889

ALIENATION AND THE METAPHYSICS OF NORMATIVITY

ON THE QUALITY OF OUR RELATIONS WITH THE WORLD

Jack Samuel

> Philosophy is to meet its need … by running together what thought has put asunder, by suppressing the differentiations of the concept, and restoring the feeling of essential being."
>
> —G.W.F. Hegel, *The Phenomenology of Spirit*

> Our picture of ourselves has become too grand, we have isolated, and identified ourselves with, an unrealistic conception of the will, we have lost the vision of a reality separate from ourselves."
>
> —Iris Murdoch, *The Sovereignty of Good*

Metaethical inquiry is at least partly a matter of making sense of ourselves, of the dimension of our lives that involves thinking and acting as moral agents. What we are doing matters to us because it is about us. I am interested in particular in two sets of potential consequences of accepting a metaethical theory: what it would mean to understand ourselves as the kinds of agents a theory envisions and what it would mean to understand our relations with one another through the theory's lens.[1]

[1] In recent years, metaethicists have along similar lines become increasingly concerned with the question of what it would mean for us if a theory of normativity were true. In contrast to conventional appeals to theoretical virtues, or to the consequences of supposedly more fundamental accounts of linguistic meaning or ontology, Parfit, for example, famously claimed that if nonnaturalism is false then nothing matters, and he and his colleagues have wasted their lives (*On What Matters*, vol. 2). Others have invoked a deep sense of angst that underlies the conviction that realism must be true (Blanchard, "Moral Realism and Philosophical Angst"), or even the first-order moral consequences of philosophers accepting realism or expressivism (Hayward, "Immoral Realism"). This approach is not entirely new—as Hayward notes, he is entering a decades-old debate between Dworkin, Blackburn, and Williams *inter alia*. My sense, however, is that these sorts of considerations have recently begun to gain traction. See also Bedke, "A Dilemma for Non-naturalists";

I argue that metaethicists should be concerned with two kinds of alienation that can result from theories of normativity: alienation between an agent and her reasons, and alienation between an agent and the concrete others with whom morality is principally concerned. A theory that cannot avoid alienation risks failing to make sense of central features of our experience of being agents, in whose lives normativity plays an important role. The twin threats of alienation establish two desiderata for theories of normativity; however, I argue that they are difficult to jointly satisfy.[2]

I begin in section 1 by saying more about what I mean by "alienation," and then, in section 2, I elaborate on what I will call the threat of *normative alienation*: that a theory of normativity could leave agents estranged from the normative facts that the theory explains. Here I draw on a few familiar literatures and argue that they express different flavors of the same underlying anxiety. In section 3, I elaborate on what I will call the threat of *social alienation*: that the normative structure of social relations envisioned by a theory of normativity would leave us estranged from one another.

The threat of normative alienation points toward a need to center the agent (the subject, the valuer, the reasoner, etc.) in a theory of normativity. The idea of "centering" the agent will, for now, have to stand as a useful metaphor, buttressed by its application to familiar examples: constructivists, subjectivists, and quasi-realists all center the agent, in the relevant sense.[3] As a first pass, the idea is that the agent is first in the order of explanation, or the order of conceptual priority. *Agent-centered* theories of normativity (typically though not necessarily antirealist) are well positioned to explain what normative facts have to do with agents, but limit themselves to bringing others into view indirectly:

and Zhao, "Meaning, Moral Realism, and the Importance of Morality." Though I will not engage directly with any of these arguments for or against metaethical positions, my aim is to establish a set of criteria motivated by a similar methodological orientation toward the theory-as-self-understanding.

[2] A theory of normativity, as I will use the term, consists in an explanation of what reasons are, and perhaps which ones there are, or of what normative facts are, and perhaps which ones are true. In what follows, I will speak interchangeably about reasons and normative facts, or about normativity in general, depending on what fits best in context. Nothing, I hope, hangs on the distinction, even if it turns out that normative facts are not in the first instance facts about reasons, *contra* the "reasons-first" orthodoxy.

[3] It is difficult to be more precise in advance of laying out the relevant features these views have in common, as I do in section 2, but see the conclusion for more elaboration. To head off one likely misunderstanding, however, I do not mean it in the sense that is roughly synonymous with "agent relative" and contrasts with "agent neutral," as in Scheffler, *The Rejection of Consequentialism*.

as a consequence of accepting universal prescriptions, or as the content of a valuing attitude, for example.

The threat of social alienation points toward a need to center the object of moral demands—the other—but the resulting *other-centered* theories of normativity (typically though not necessarily realist) will have difficulty accounting for the significance of normative facts to agents. Metaethical accounts suited to accommodate the role of others in our normative lives ground normativity in, e.g., facts about concrete others, or the relations we stand in to them. But facts about our relationships to others, or the properties possessed by others, are not the right sorts of facts to ensure that we will have the right kind of connection to them.

If this is right, a theory of normativity suited to avoid both forms of alienation would paradoxically seem to need to center both the agent and the other. The tension can be resolved, however, by centering the constitutive relations between agents as such and others. To paraphrase Michael Thompson, metaethics must be able to record the special sort of dent that *others* themselves make on one's *own* agency, on pain of leaving us in one state of alienation or another.[4]

1. Alienation in General

A natural worry that is worth addressing before I begin is that, without some account of what alienation is, organizing the following problems under that heading will have diminished explanatory potential.[5] It is not, after all, a stable or uncontested concept. In the most general use it is more or less synonymous with "separation," as in the "alienation" of property rights through contract. Philosophers tend to use the term with a negative valence, as synonymous with estrangement—making strange. While alienating one's property rights through contract is putatively neutral or even good, being alienated from the products of one's labor, from nature, or from God is bad. An alienated relationship with something is a defective form of that relationship. In its most general form, alienation is a problematic separation between a person (a subject, an agent) and something else, something from which we ought not to be separate.

Alienation and the critique thereof operate on a number of levels. In the first instance, alienation may be a feature of a way of life or a mode of social organization, as when capitalism allegedly alienates workers from the products of their labor. But insofar as this kind of alienation is subject to critique (and not just material social or political intervention) it is because the alienated

[4] Thompson, "What Is It to Wrong Someone?," 346.

[5] Thanks to an anonymous reviewer for encouraging me to address this worry.

mode of social organization embodies an alienated conception of ourselves. We can thus critique in philosophy the underlying picture of the human, the person, the worker, as a way of making explicit the distortion of social organization it produces or reflects. This sort of critique appears in the work and interpretations of "continental" figures like Hegel, Marx, Lukács, Heidegger, and Fromm.[6] Capitalism may be (or at least require) a defective relationship between a person as producer and the product of their labor, and is thus a defective form of social reproduction, which embodies a defective picture of the nature of human agency.[7]

At one level of abstraction higher, but in more or less the same tradition, we might say that a *theory* alienates us insofar as it tends to lead to our living alienated lives if we adopt it, or if it informs the cultural backdrop against which we live. On the other hand, we might say that a theory is itself alienating, or embodies alienation, insofar as it represents agents such that if we *were* the way the theory envisions us then we would be alienated, or insofar as it obscures, *qua* theory, that from which we risk being alienated. This use is probably more familiar in Anglophone philosophy, where worries about alienation are often associated with Bernard Williams or Peter Railton.[8]

In my view, however, they come to the same thing: the alienation at issue is between a person and something from which persons are not properly separate, and it can be realized in a social relation, a mode of production, a theory that informs a social relation or mode of production, or a theory that holds itself out as giving us some insight into what kinds of things we are. Where there is in human life—the life of the metaethicist, and of those they imagine as their subjects of inquiry, for my purposes—a harmony or unity or cohesion or familiarity, an alienating theory imagines us as held apart from that with which we are in reality united and familiar. It makes those things strange to us, and if we could manage to truly understand ourselves as the theory encourages us to, we would suddenly be puzzled by the commonplace, unable to make sense of some important part of our own lives. That is the sense in which, as I will argue,

[6] For survey and reconstruction, see Schacht, *Alienation*; and Jaeggi, *Alienation*.

[7] For an interpretation of Marx along these lines, see Julius, "Suppose We Had Produced as Humans." There is some reconstruction involved in attributing to Marx a concern for correctly conceiving of human agency, but for the sake of assimilating the Marxian critique of alienation into a larger story about the concept, I trust that it is sufficiently well founded. See also Honneth, "Foreword": "The concept of alienation . . . presupposes, for Rousseau no less than for Marx and his heirs, *a conception of the human essence*: whatever is diagnosed as alienated must have become distanced from, and hence alien to, something that counts as the human being's true nature or essence" (vii, emphasis added).

[8] See sections 2.2 and 3.1 below.

a theory of normativity threatens to *alienate* us from it, by casting it as a strange and distant thing, rather than as something that suffuses or partly constitutes our experience of ourselves and others.

2. Normative Alienation

If a normative theory is to offer a satisfying account of reasons it must be able not only to tell us what reasons are, and perhaps which ones there are, but what they have to do with us. It must be able to explain normative facts in a way that connects them to the individuals they are normative for in the right way. In doing so, it will avoid normative alienation.

A normatively alienated agent would be one for whom normative facts were recognizably true, but irrelevant or obscure. They would be, so to speak, *mere* facts, like the fact of whether or not Golbach's conjecture is true, or the fact of how many stars there are in a distant galaxy: suitable objects of curiosity but possibly unknowable, of no consequence to us in our ordinary lives, or both.

Moral facts cannot be facts like these, and this image of agency—mere receptivity to such facts—cannot represent ours. The first desideratum for a theory of normativity is that in its explanation of how normative facts can be true it contains an explanation of how they are normative *for us*.[9]

The threat of normative alienation appears in different guises: that normative facts could fail to be motivating, that they could fail to be acknowledged as authoritative, and that they could fail to be identifiable. Each of these concerns corresponds to a familiar debate in recent metaethics but they are generally not recognized as expressions of a more general anxiety.[10] One thing that they do have in common, however, is that they underlie many of the familiar challenges to traditional forms of normative realism and are offered in support of various agent-centered alternatives.[11] This is, I argue, no accident. Traditional

[9] This is, in a way, Kant's demand to explain how reason can be *practical*—see, e.g., *Groundwork of the Metaphysics of Morals*, 4:395, 448, and *Critique of Practical Reason*, 5:444–46.

[10] Shamik Dasgupta identifies the first and second guises of normative alienation as versions of the same desideratum, though he does not include the epistemic challenge or characterize them as a threat to properly conceiving of normativity's relation to agents. See Dasgupta, "Normative Non-naturalism and the Problem of Authority."

[11] "Agent centered" and "realist" are not antonyms in my usage. Mark Schroeder's Humeanism is a form of reductive realism about reasons that centers the desires of the agent in its explanation of what reasons there are and which ones exist. See Schroeder, *Slaves of the Passions*. Metaethical constructivism, Kantian (e.g., Korsgaard, *The Sources of Normativity* and *Self-Constitution*) and Humean (e.g., Street, "Constructivism about Reasons"), is a paradigmatically agent-centered approach to metaethics, and is sometimes characterized as a form of *procedural* realism about normativity. Agent-centered metaethics contrasts

forms of moral realism threaten to vindicate the truth of normative facts at the expense of undermining the intimacy of their connection to agents. Realists themselves are of course typically untroubled by this, but many (most?) of us find the idea intolerable. We find it intolerable in different ways, and it is not always clear that it is the same thing we find intolerable. But, I argue, these debates represent more local ways in which philosophers have struggled to bring normativity closer to us, and thus have a common source in an implicit concern for something like normative alienation.

If it were possible for us to be alienated from morality in the way that this anxiety concerns, morality would not be fit to play the role in our lives that it evidently does. The truth, reality, or objectivity of normative facts would have been purchased at the cost of their relevance.

2.1. The Constructivist Challenge: Normative "Grip"

It is common to characterize Kantian constructivism as an attempt to avoid naturalistic objections to traditional realism without losing the objectivity of moral talk (as noncognitivism is often thought to do).[12] But it is in my view Korsgaard's

rather with what I will sometimes call "traditional" forms of realism: nonnaturalist (e.g., Moore, *Principia Ethica*; Ross, *The Right and the Good*; Scanlon, *What We Owe to Each Other*; Parfit, *On What Matters*, vol. 1; Enoch, *Taking Morality Seriously*; and Shafer-Landau, *Moral Realism*) and naturalist (e.g., Railton, "Moral Realism"; Boyd, "How to Be a Moral Realist"; Brink, "Externalist Moral Realism"; and Sturgeon, "Moral Explanation"), wherein the truth of normative facts is explanatorily independent of the agents for whom they are normative, and they become practical for agents only by being discovered (and perhaps further by being discovered in relation to the agent's desires).

[12] In the last decade, the conventional wisdom has consolidated around the idea that what speaks in favor of metaethical constructivism, if anything, is its ability to balance a handful of theoretical desiderata. Facing a stalemate between realism and antirealism, constructivism supposedly aims to recover the *objectivity* of moral facts from the prevailing noncognitivism of the mid-twentieth century, and to do so without running afoul of the *naturalistic* worries associated with critics of traditional (intuitionist) moral realism (e.g., Mackie, *Ethics*). What exactly objectivity comes to is a matter of dispute, but it is something like that there are normative facts, or facts about reasons, irrespective of what anyone in particular thinks; that our normative judgments or judgments about reasons are truth apt and at least sometimes true; or that genuine cognitive disagreement about normative facts or facts about reasons is possible. Thus constructivism splits the difference, rendering moral facts genuinely objective while naturalistically respectable.

In other words, constructivism offers a way of being a naturalist (which we all want in a post-Mackie world) *and* a cognitivist (which we all supposedly want in a post Frege-Geach world), something thought difficult to pull off before the Tanner Lectures that became *The Sources of Normativity*. Or at any rate, this, I take it, is the received view of what the problem is for which constructivism is supposed to be a solution. Enoch summarizes this motivation for the Kantian constructivist nicely:

key insight that metaethics must avoid what I am calling normative alienation.[13] She argues that traditional realism leaves an explanatory gap between normative facts and the agents for whom they are normative—that realists hold that "we have normative concepts because we have spotted some normative entities, as it were wafting by."[14] In other words, even if (*contra* Mackie) there were entities answering to the realist's needs it would be a mistake to understand moral language as merely registering their existence, rather than having an essentially practical role. If such entities are just sitting there among the furniture of the universe it would be mysterious how they could get a grip on us—address us *as agents*—how they could feature centrally in the exercise of practical reason. Constructivism proposes to explain normative facts in a way that connects them to the individuals they are normative for in the right way.[15]

The task for constructivism is thus to account for the non-accidental harmony of reasons for action and our capacity to act for reasons. It begins by

> Many people are suspicious about more robust, non-procedural forms of metanormative realism. They think that there are serious metaphysical and epistemological worries (and perhaps others as well) that make such realism highly implausible. Nevertheless, going shamelessly antirealist also has problems. We seem to be rather strongly committed, for instance, to there being correct and incorrect ways of answering moral (and more generally normative) questions, and moreover our moral (and more generally normative) discourse purports to be rather strongly objective. Constructivism may be thought of as a way of securing goods realism (purportedly) delivers, for a more attractive price. (Enoch, "Can There Be a Global, Interesting, Coherent Constructivism about Practical Reason?" 324)

The metaphysical worries that Enoch gestures toward here are associated with "Mackie's problem." They express the suspicion that there could be entities answering to the traditional realist's needs. This is *a* problem for which constructivism might provide an answer, but representing the dialectic this way ignores the internal motivation that Korsgaard offers.

In addition to Enoch, for this understanding of what motivates Kantian constructivism, see Tiffany, "How Kantian Must Kantian Constructivists Be?"; Shafer-Landau, *Moral Realism*; Lenman and Shemmer, "Introduction"; and arguably Smith, "Search for the Source"; and Gibbard, "Morality as Consistency in Living." In fairness, Korsgaard does cite something like naturalistic scruples as motivation in the prologue to *The Sources of Normativity* ("The ethics of autonomy is the only one consistent with the metaphysics of the modern world," 5), but on my reading of Korsgaard, this is not the central question for which constructivism is supposed to be an answer.

[13] See Samuel, "Toward a Post-Kantian Constructivism," sec. 1.

[14] Korsgaard, *The Sources of Normativity*, 44.

[15] Bagnoli makes a similar point in arguing that the "standard objection" to Kantian constructivism rests on a mistaken understanding of its basic claim to explain the bindingness of reasons in terms of the activity of reason*ing* (see "Kantian Constructivism and the Moral Problem").

acknowledging that the demand to understand reasons arises in the first place out of the fact that insofar as we occupy the practical standpoint we rely on them:

> Normative concepts exist because human beings have normative problems. And we have normative problems because we are self-conscious rational animals, capable of reflection about what we ought to believe and to do.[16]

As Scanlon summarizes the worry on Korsgaard's behalf (though he is unpersuaded), "If a consideration's being a reason for a person is just another fact about the world ... then the person could still be perfectly indifferent to this fact."[17] The worry is that simply ascribing to certain facts a very special kind of property leaves mysterious why it should appear in our deliberation:

> There are certain things that we ought to do and to want simply because they have the normative property that we ought to do or to want them (or perhaps I should say that they ought to be done or to be wanted). The synthesis between the oughtness and the action, or the agent and the oughtness—however that is supposed to go—cannot be explained. It is like a brute fact, except that it is at the same time an a priori and necessary fact.[18]

[16] Korsgaard, *The Sources of Normativity*, 46.

[17] Scanlon, *Being Realistic about Reasons*, 9.

[18] Korsgaard, "Normativity, Necessity, and the Synthetic *A Priori*, 2; cf. Korsgaard, "Realism and Constructivism in Twentieth-Century Moral Philosophy":

> If it is just a fact that a certain action would be good, a fact that you might or might not apply in deliberation, then it seems to be an open question whether you should apply it. The model of applied knowledge does not correctly capture the relation between the normative standards to which action is subject and the deliberative process. And moral realism conceives ethics on the model of applied knowledge. (112)

Here Korsgaard follows Rawls, for whom constructivism is explicitly an approach to theorizing normativity that begins with the difficulty of finding a way to live together—an essentially practical project—rather than with the theoretical investigation of a special kind of truth: "The search for reasonable grounds for reaching agreement rooted in our conception of ourselves and in our relation to society replaces the search for moral truth interpreted as fixed by a prior and independent order of objects and relations, whether natural or divine, an order apart and distinct from how we conceive of ourselves" ("Kantian Constructivism in Moral Theory," 519). Realists like Scanlon and Parfit resist the idea that there is anything to be explained: it simply is the nature of the property of rightness, goodness, oughtness, or being a reason that insofar as we have the capacity for practical reason any bearer of the property is a fitting object for its exercise. As Scanlon puts it, "It seems to me that no such further explanation of reasons need or can be given: the 'grip'

The idea at the core of Korsgaard's project is that metaethics will leave us alienated from normativity if metaethics does not offer an explanation for normativity's connection to agents. Her solution is to center the agent, understood in terms of the reflective capacity to act for reasons, in the explanation of how there can be normative facts.

2.2. The Humean Challenge: Motivation

Perhaps the most familiar expression of anxiety about normative alienation, though it does not present itself in these terms, is the "Humean" challenge to motivational externalism about reasons. "Internal reasons theorists" hold that it is a necessary condition on something's being a reason for an agent that it stands in some relation to motivational facts about her. Exactly what relation and exactly what kind of motivational facts vary, but the underlying thought is that if it is not *possible* (for some sense of possibility) for an agent to be motivated by something then it cannot be a reason for her.

Internal reasons theorists do not generally frame their position in terms of avoiding alienation. Insofar as Hume held a view like this it followed from his more basic metaphysical commitments, and in the recent literature internalism is sometimes framed as an analysis of the concept of a reason or of reasons talk, where it is part of the very idea of something's being a reason that it is related to one's motivations in a certain way.[19] However, I suspect that the enduring appeal of the position depends at least in part on anxieties (explicit or implicit) about alienation: if there were "external reasons" then they could fail to be motivating, but reasons must be capable of motivating us, so there could not be external reasons. In other words, external reasons, if there were any, would be distant from us in a way that they could not be while still playing the role that we take them to in our lives. Railton glosses the basic idea similarly, bringing out the dimension of this debate that corresponds to what I am calling normative alienation:

> Absent a link between moral judgment and motivation, ethics might as well be speculative metaphysics. What else could account for the distinctive way in which moral judgments are normative—"action guiding"—for the agent who makes them?[20]

that a consideration that is a reason has on a person for whom it is a reason is just being a reason for him or her" (*Being Realistic about Reasons*, 44).

[19] On Hume's metaphysical internalism, see Schafer, "Hume on Practical Reason." On internalism as an analysis of the concept of a reason, see Finlay, "Responding to Normativity" and "The Obscurity of Internal Reasons."

[20] Railton, "Internalism for Externalists," 171.

This interpretation of the impulse underlying the Humean challenge finds support in Williams's inaugural contribution to the debate.[21] There he argues against the possibility of external reasons on the basis that if there were any they could not motivate us. He accepts that external reasons correspond to something in ordinary language but denies that there could be any because they would be unfit to play an explanatory role that he thinks reasons must: "If something can be a reason for action, then it could be someone's reason for acting on a particular occasion, and it would then figure in an explanation of that action."[22] That there could not be external reasons because, if there were, they could not enter into the explanations of agents' actions is plausibly an expression of an anxiety about normative alienation: if there were any external reasons, they would be (at least sometimes) irrelevant to us, and this cannot be.

Read in the context of Williams's larger body of work, this interpretation gains further plausibility. One of Williams's persistent concerns is to vindicate a nonalienated conception of agency. What this amounts to for him is that as agents we are defined by our projects, values, and commitments, in a way allegedly threatened by utilitarian and Kantian moral theory.[23] His work is animated by the conviction that things are going wrong if we conceive morality as the business of some isolable, rational part of the soul, whose task is to discover what reasons there are out there in the universe.

In the iconic "one thought too many" thought experiment, he notes that an agent who reasoned that it was permissible to save their drowning spouse over a stranger will have already gone wrong in posing the question, rather than being moved directly by the recognition that it is their own spouse. To think that settling the question of what to do requires transcending the embodied particularity of oneself as an actual agent, in search of facts commanding objectivity or universality, is to lose one's grip on oneself.

At the level of moral theory, Williams insists on bringing ethics "closer" to the agent, preserving an intimate connection between who we are as distinctive agents and what we have reason to do, even if it means opting for a moderate form of moral nihilism. In this connection, his denial that there could be reasons that fail to enter into the explanations of agents' actions appears to be part of a larger effort that cuts across the putative distinction between moral theory and metaethics: an effort to make normativity *human*, to restore its connection to us.

[21] Williams, "Internal and External Reasons."

[22] Williams, "Internal and External Reasons," 106.

[23] See Williams, "A Critique of Utilitarianism," "Moral Luck," and "Persons, Character, and Morality."

It should not be controversial that avoiding alienation by humanizing moral theory is a persistent concern for Williams. I hope that I have made it plausible that he is concerned with a form of alienation not only where he explicitly invokes it as a problem for moral theory but in his moral psychology—that is, that at least for Williams reasons internalism is a part of his campaign to avoid alienation. This does not prove that the Humean challenge in general is really about avoiding alienation: there may be some internal-reasons theorists for whom avoiding normative alienation is at most a welcome but unimportant subsidiary benefit. Nevertheless, the Humean challenge *can* be understood as an expression of an anxiety about alienation, and it is this connection to a deep philosophical impulse, more than technical problems about the analysis of language, that I suspect explains its perennial appeal. Insofar as metaethics is, as I have suggested, in the business of helping us make sense of ourselves, it makes sense to worry that external reasons, if there were any, would be troublingly disconnected from our lives.

2.3. The Epistemic Challenge

Probably the least remarked-on guise of normative alienation is its epistemic one. A theory of normativity that vindicated the truth of normative facts but allowed that they were epistemically distant from us would leave us intolerably estranged from them. It is sometimes claimed that normative facts must be knowable for agents *in virtue of being agents*, that there must be a "non-accidental connection between the normative truth and our faculties for forming normative beliefs."[24] Less controversially, we need some explanation for the knowability of normative facts in order not to be epistemically alienated from them. As Thomas Nagel, himself a realist, puts it:

> The connection between objectivity and truth is therefore closer in ethics than it is in science. I do not believe that the truth about how we should live could extend radically beyond any capacity we might have to discover it (apart from its dependence on nonevaluative facts we might be unable to discover).[25]

This generates a familiar challenge to traditional realists—namely, that they can offer no explanation for why, if the truth about how we should live is simply out there, this knowledge is possible. Like most realists, Nagel is content not to offer one, but advocates of agent-centered approaches to metaethics generally—and constructivists in particular—tend to emphasize not only that we

[24] Schafer, "Realism and Constructivism," 709.

[25] Nagel, *The View from Nowhere*, 139.

should want such an explanation but that there are special obstacles realists face to offering one.[26]

In her classic argument against realism and in favor of Humean constructivism, Sharon Street, for example, appeals to the knowability of normative facts as something that realism cannot explain.[27] If normative facts were radically mind independent it would be at best a matter of luck that we were able to track them with our normative judgments. Street relies on the perhaps controversial premise that humans come by our evaluative attitudes largely as a result of evolutionary forces, but the claim can be stated more generally: presumably insofar as we are natural creatures our evaluative attitudes are susceptible to empirical explanation, and such explanation will be independent of the truth of the corresponding normative facts. Thus, realists must be able to explain the relationship between whatever causal forces such empirical explanations invoke (evolutionary psychological or otherwise) and the truth of the relevant normative facts: a challenge that Street argues no realist can meet.

Street's own view, Humean constructivism, holds that normative facts are determined for each agent by her own normative judgments, and thus are knowable through the activity of making them explicit and bringing them into coherence. Kantian constructivism as well can boast a ready explanation for their knowability for agents as such: that it is the exercise of practical reason that determines them.

Constructivists are not the only ones to press this challenge. Mark Schroeder notes that "irrealists" of different stripes can easily account for normative knowledge, and that reductivists in particular take this to speak in their favor. Given that realists find it especially difficult to do so, Schroeder notes that "the main divide among realists between reductivists and non-reductivists used to be characterized as the dispute about whether intuitionism is true."[28] In other words, the fate of non-reductive realism depends on realists' ability to defend their rejection of having to explain the possibility of moral knowledge, over and above merely asserting it. As with the challenge to explain normativity's "grip" on us as agents, traditional realists tend to respond to the puzzle of how moral knowledge is possible simply by claiming that it is. Or anyway, this is how anti-realists and reductive realists tend to see things.

[26] For a discussion of this point, see Schafer, "Realism and Constructivism."

[27] Street, "A Darwinian Dilemma for Realist Theories of Value."

[28] Schroeder, *Slaves of the Passions*, 170; cf. Harman, *The Nature of Morality*. Schroeder is relying on a different taxonomy of metaethical theories, but in contrast to non-reductive realism, against which he presses a version of the epistemic alienation worry, the views he identifies as incurring no special epistemic burden are agent centered in my terms.

For those who find it mysterious or even occult that we should see normative facts as essentially knowable, without further explanation, concerns about moral epistemology put pressure on approaches to metaethical theory that do not center the agent as the bearer of practical knowledge. To accept a traditional realist account of the explanation of normative facts while remaining skeptical of the realist's non-explanation of their knowability would leave one in a state of epistemic alienation, convinced that there were normative facts but with no way of discovering what they were.

2.4. The Solution: Agent-Centered Metaethics

These classic objections to realism express related anxieties about the possibility that we could have reasons to which we were motivationally indifferent, reasons whose relevance to our activity of reflective self-determination was at best coincidental, or reasons of which we could in principle be systematically unaware. If it were possible for reasons to be like that, they would be totally estranged from us. These more local challenges to traditional moral realism are thus expressions of a sense that morality cannot be alien to us, and that a theory of normativity must come along with an explanation of how it can be *ours*.

The threat of normative alienation calls for a theory of normativity that brings it closer to us, intermingling it with the messy, embodied, and perhaps contingent features of human life with which we each individually have the most direct familiarity. The resulting proposals all center the individual agent in their derivation of normativity, emphasizing desires, values, preferences, or the embodied capacity to practically self-determine, as in some sense foundational to the explanation of how there could be such a thing as normativity at all. In the next section, however, we will see that in bringing normativity closer to ourselves we risk losing our moral grip on one another.

3. SOCIAL ALIENATION

In section 2, I argued that several familiar challenges to traditional metaethical realism can be understood as expressions of a more general underlying anxiety, an anxiety about the possibility that morality could be alien to us. A theory of normativity that failed to grapple with this fact would fail to capture something important about the experience of being a moral agent. Though not everyone is moved by all or even any of these challenges, I take it that I can help myself at least to their plausibility.

In this section, however, I will raise a different kind of challenge, one that reflects a different kind of anxiety: that moral theory might represent us to ourselves as estranged from one another. Corresponding to this anxiety is the second desideratum for a theory of normativity: to explain how it can be that we are morally related to concrete others, and thus to avoid what I call *social alienation*.

This desideratum has gone largely unrecognized and is difficult to formulate using ready-to-hand conceptual resources.[29] As a first pass, the challenge is to ground the essential sociality of morality.[30] Much of morality involves responding to the grip we have on each other. Agent-centered theories run the risk of erasing this distinctive grip, making agency out to be a matter of individuals following rules, recognizing reasons, or standing in relation to themselves (the relation of reflective distance, for example), giving us at best an indirect way to recognize other people. To begin to bring this worry into view, I return to Williams.

3.1. Alienation in Twentieth-Century Moral Theory

A persistent theme in Williams's work is that ethics must account for the ways that we are shaped as distinctive agents by our projects, commitments, and values. To the extent that moral theory alienates us from these parts of our lives, it presents an image of the moral agent in which we cannot recognize ourselves. However, while the examples that Williams uses to motivate his objections typically feature important social relationships, his diagnosis of alienation interiorizes the problem, making it an individual, psychological defect, and not a social one.

Utilitarianism, for example, is a threat to an agent's integrity because "it can make only the most superficial sense of human desire and action," and it "alienates one from one's moral feelings."[31] What goes wrong in the "one thought too many" case is that the husband appeals to an explicit deontic order, thinking a judgment about duty or rules is a necessary intermediary between his affection and how he ought to act. Moral theory, he worries, "treat[s] persons in

[29] The concern has gone largely unrecognized, but not entirely. Aside from Iris Murdoch, whom I discuss in what follows, some others I think are onto something like this worry include Kate Manne, "On Being Social in Metaethics" and "Locating Morality"; Michael Thompson, "What Is It to Wrong Someone?" and "You and I"; Kenneth Walden, "Laws of Nature, Laws of Freedom, and the Social Construction of Normativity," "Mores and Morals," "Morality, Agency, and Other People," and "Reason and Respect"; and Kieran Setiya, "Other People." As in the previous section, none of my antecedents have explicitly identified social alienation as something to be avoided, but I think their interventions can be profitably understood, along the same lines as mine, as taking the sociality of morality seriously in a way that has metaethical implications.

[30] Social alienation is a problem for morality specifically. It may turn out in the end that the best theory of normativity implies that all normativity is social; cf. Brandom, who argues that the normativity of meaning is social (*Making It Explicit*). But it is not a demand on a theory of normativity that it explain the sociality of all normativity, only that it explain normativity in general in a way that does not rule out the essential sociality of morality.

[31] Williams, "Critique," 82, 104.

abstraction from character," making us out to be nothing more than a "locus of causal intervention in the world."[32]

The "one thought too many" case highlights a disconnect between moral theory and human life, realized in an agent's (in)ability to properly understand himself in relation to another. However, Williams's understanding of alienation and integrity points toward achieving internal, psychological unity (something like virtue) as the solution.

The contrast comes out more clearly in the work of two contemporary critics of alienation in moral theory: Michael Stocker and Peter Railton. Stocker's central case involves someone explaining their choice to visit a friend in the hospital by appealing to the duties of camaraderie, and Railton responds to a worry about someone regarding their spouse as a mere source of utility. For Stocker, "moral schizophrenia" consists in a disunity of one's motivations and values.[33] "One mark of a good life," he claims, "is a harmony between one's motives and one's reasons, values, justifications."[34] If moral theory is to help us understand what it is to live a good life, it must be able to make sense of how such harmony is possible. For Railton alienation involves our affective selves coming apart from our rational, deliberative selves: "there would seem to be an estrangement between [an agent's] affections and their rational, deliberative selves; an abstract and universalizing point of view mediates their responses to others and to their own sentiments."[35] Both critiques are motivated by noting a defective form of sociality, allegedly due to adopting an alienating moral theory, and both diagnoses identify psychological disunity as the problem, and psychological unity as the solution.

Unlike Williams and Stocker, Railton hints at something like the problem of social alienation as I conceive of it—estrangement between oneself and another—as an equally important dimension along which moral theory can be alienating, and one from which the psychological is not cleanly separable. He notes that "we should not think of John's alienation from his affections and his alienation from Anne as wholly independent phenomena, the one the cause of the other."[36]

In establishing the criteria for an adequate response to the problem of alienation, he emphasizes the role that relationships with others must be allowed to play:

[32] Williams, "Moral Luck," 19, and "Critique," 96.

[33] Stocker, "The Schizophrenia of Modern Ethical Theories."

[34] Stocker, "The Schizophrenia of Modern Ethical Theories," 453.

[35] Railton, "Alienation, Consequentialism, and the Demands of Morality," 137.

[36] Railton, "Alienation, Consequentialism, and the Demands of Morality," 138.

> First, we must somehow give an account of practical reasoning that does not merely multiply points of view and divide the self—a more unified account is needed. Second, we must recognize that loving relationships, friendships, group loyalties, and spontaneous actions are among the most important contributors to whatever it is that makes life worthwhile; any moral theory deserving serious consideration must itself give them serious consideration.[37]

He warns against "the picture of a hypothetical, presocial individual" by which philosophers have become distracted, which results in unthinkingly assuming that self-concern is natural and requires no special explanation, while concern for others is taken to require one. A solution, he suggests, must capture the importance of "participation in certain sorts of social relations—in fact, relations in which various kinds of alienation have been minimized," and that the starting point for moral theory must be the "situated rather than presocial individual."[38]

However, Railton ultimately leaves the problem under-theorized. If there is a social dimension to these cases that has been mostly ignored, what demand does it place on the theorist? Here I only have the space to offer a sketch of a view that I elaborate on elsewhere.[39] The key upshot is that avoiding social alienation—achieving social integrity, to repurpose Williams's distinction—requires that in our ethical self-awareness we account for the significance for us as agents of others as *external*, as *particular*, and as *subjects*—as each an individual reality, separate from oneself.[40] We must be able to make sense of ourselves, that is, as responsive to others themselves, not just to rules for conduct that make reference to others in their application conditions; to particular others,

[37] Railton, "Alienation, Consequentialism, and the Demands of Morality," 139.

[38] Railton, "Alienation, Consequentialism, and the Demands of Morality," 168, 147, and 171.

[39] For the source from which the above line of exegesis is drawn, see Samuel, "An Individual Reality, Separate from Oneself."

[40] We can see the distinction more clearly by reflecting on an analogous puzzle about the epistemology of perception, concerning how we can have perceptual experience of the world itself and not merely of our inner representations of it. Not everyone agrees that this is something to be achieved, but those that are concerned with the threat of being trapped behind the "veil of ideas" (perceptual alienation from the world) tend to emphasize both externality and particularity as important features of worldly objects *qua* worldly. See Brewer, *Perception and Reason*; Martin, "On Being Alienated"; Travis, "The Silence of the Senses"; and McDowell, "Criteria, Defeasibility, and Knowledge." The phrase "individual reality, separate from oneself" is a patchwork of two different phrases Murdoch uses in *The Sovereignty of Good*: her gloss on Simone Weil's concept of attention as a "just and loving gaze directed upon an individual reality" (33) and her characterization of the object of moral awareness as "a reality separate from ourselves" (46).

not just to abstract idealizations of others as representative rational agents, persons, and so on; and to others as subjects, and thus potentially responsive to us. I refer to a form of moral self-awareness that meets these conditions as the achievement of "practical openness to the other."[41] My practical openness to another is not separable from the other being practically open to me—otherwise we would each only be open to one another as to a third person that we each see as bearing a special normative property, rather than as standing in relation to ourselves.

Integrity, for Williams, is a matter of an agent's moral thought and action staying close to everything else that makes her her. Social integrity, as I have been sketching it, is a matter of one's moral thought and action reflecting mutual practical openness to others. If socially alienated moral knowledge is the mere apprehension of one's reasons or the rules by which one is bound, socially integrated moral knowledge is an awareness of others as such. The threat of social alienation in ethics is a kind of normative solipsism. To avoid social alienation is to account for what Iris Murdoch characterizes as "the extremely difficult realization that something other than oneself is real."[42]

The phenomenon of practical openness to the other is in my view tragically undertheorized, and this is not the place to attempt a project of that scope. With a hazy idea of the problem in view, in order to give a sense of the stakes I will offer an example of how it manifests in a set of issues in normative ethics: the phenomenon of "directedness." Recognizing another as the object of a directed obligation is a case of practical openness to another, and one a proper understanding of which is threatened by agent-centered metaethics.

3.2. *Directedness in Ethics*

An obligation is "directed" when it is owed to someone in particular. Perhaps we are all obligated to give to charity, but we do not owe it to any particular charity to give to them. We are also obligated to keep our promises, but in each case we owe it to the promisee. Directed obligations are generally thought to correlate with or be identical to claim rights, so another way to put the point would be that no particular charity has a claim on our beneficence, but each time we make a promise we grant to the recipient a claim to our performance. When we violate a directed obligation we do not merely do something wrong but *wrong* someone in particular: the one to whom the obligation is owed. The

[41] Samuel, "An Individual Reality, Separate from Oneself," 14, paraphrasing John McDowell's slogan that avoiding what I called perceptual alienation requires epistemic "openness to the world"; see McDowell, *Mind and World*.

[42] Murdoch, *The Sublime and the Good*, 215.

one who is wronged is thus in an important sense the *victim*, not merely the *occasion* of wrongdoing.[43]

Directed obligations constitute the core of morality. They reflect what Wallace calls the "moral nexus" that joins concrete persons, equally real.[44] Being aware of and responsive to standing to others in a moral nexus is an important way, if not *the* fundamental way, of being practically open to one another. The moral nexus is a basic social relation that arguably cannot be explained in terms of reasons, rules, and putatively more normatively fundamental self-relations. A metaethics without the resources to capture the moral nexus risks theorizing away the sociality of morality.

One way a metaethics might run this risk would be by purporting to directly entail a normative ethical theory with no room for directed obligations at all (say, act consequentialism). More subtly, a metaethics might entail that directed obligations are not really directed. Along these lines, Aleksy Tarasenko-Struc argues that Korsgaard is committed to the view that obligations apparently owed to others are in fact owed to ourselves. Because Korsgaard grounds all normative authority in the constitutive ability of agents to bind themselves, he argues that all obligations are ultimately grounded in this self-relation: "The problem is that she embraces an egocentric conception of authority, on which we originally have the authority to obligate ourselves whereas others only have the authority to obligate us because we grant it to them."[45] There will always be an unbridgeable explanatory gap between obligations to oneself and those apparently owed to another.

From the fact that Korsgaard grounds obligations to others in obligations to oneself it does not obviously follow that obligations to others are illusory. They would be derivative, but a derivative obligation may bind all the same. The worry is that it may not bind *in the right way*—that is, that an obligation that derives ultimately from the individual requirements of self-constitution will turn out not to be a genuine instance of being bound by another, but only appear so. The explanatory challenge for Korsgaard is to explain how an obligation that derives from an obligation to oneself will not turn out, on careful inspection, to be merely an obligation *to* oneself that *concerns* another, depending on how the derivation is fleshed out.

If I make a promise to myself to smile at strangers more, the promise becomes concerned with a stranger when he walks by because he is an opportunity for

[43] This way to refer to the distinction is due to Thompson, "What Is It to Wrong Someone?," 340.

[44] Wallace, *The Moral Nexus*; cf. Nagel, *The Possibility of Altruism*.

[45] Tarasenko-Struc, "Kantian Constructivism and the Authority of Others," 77.

me to keep the promise. My smile, however, is not owed to him. If I am in a bad mood, I do not wrong him by maintaining a neutral expression, unless we make eye contact and he smiles at me and I am now being rude. I act wrongly *vis-à-vis my promise to myself* in my conduct *concerning* him. An obligation genuinely owed to another is not like this: it is an opportunity to do right by another or to wrong them, not just to do right or wrong.

One way Korsgaard might try to get around this problem is to hold that our authority over ourselves can be transmitted to others. On this view, I can have directed obligations to others because other people can exercise the power, which I have transmitted to them, to bind my agency.[46] In other words, rather than exercising my ability to obligate myself by binding myself to do something concerning another (smiling at strangers), I could somehow *transfer* that authority to another, to be exercised by them, thereby obligating me. It is not clear that the idea of such a voluntary transfer of authority can work. The trouble is not that authority can never be genuinely transferred: if one party with authority—say, the president—appoints an official to oversee the activity of a third party, the third party will for all practical purposes answer directly to the official. One could argue that there remains a sense in which the third party ultimately is obligated only to the president, with an official as a normative intermediary, but there is surely a recognizable sense in which the official's orders obligate the third party directly.[47] However, if the president appoints an official to oversee *himself*, on the authority of his own office, he can only ever *appear* to obey the official, for the moment the official issues an apparent command the president does not wish to follow, he can simply withdraw the grant of authority, proving the transfer to have been illusory all along.[48]

[46] See Korsgaard, *Self-Constitution*, 189–91; cf. Tarasenko-Struc, "Kantian Constructivism and the Authority of Others," 85–87. Another strategy Korsgaard could pursue, but appears not to, would be to invoke the distinction between the content and justification of a norm, like promise keeping. I address this approach as a general matter in section 3.3 below.

[47] Tarasenko-Struc makes a similar distinction between *discretionary* and *original* authority, and notes that for discretionary authority to be genuine authority it must presuppose a prior grant of original authority, which again Korsgaard cannot explain ("Kantian Constructivism and the Authority of Others," 85–86). The following argument runs parallel to his, though in slightly different terms.

While the official's orders plausibly obligate the third party directly, it does not follow that the third party owes performance to the official—see the discussion of the example of private law enforcement in section 3.3 below—but my point here is that even if we assume that in a trilateral case we get something approximating genuine transfer of authority we face a special difficulty where the original source of authority is the one putatively obligated by that same authority once transferred.

[48] Compare United States v. Nixon, 418 US 683, 706–7 (1974), holding that the president cannot be permitted to determine the extent of his own executive privilege vis-à-vis a

What is needed in the special case of voluntarily transferring one's own authority over oneself to another is some way to ensure that, once transferred, the authority cannot be voluntarily withdrawn. If we model the transfer of authority on the idea of a promise to oneself to obey another we will not get that, since it is characteristic of promises that the promisee has the ability to release the promisor (this is why the idea of a promise to oneself is suspicious to begin with). But if we can find a different model on which a power to obligate oneself can be transferred, such that when done it cannot be voluntarily undone, we will still have to confront the worry that whatever it is that prevents it from being withdrawn will require an independent source of authority, one that finds no place in Korsgaard's theory.

Supposing, however, that a genuine, voluntary transfer of authority is possible on Korsgaard's account, it will leave us with an unsatisfying asymmetry: that others have only as much authority over us as we grant them is not much of an improvement over having obligations concerning others but owed to oneself. As Tarasenko-Struc concludes, with an analogy to the classic "problem of other minds": "just as a person's wince might be thought to directly reveal that she is in pain, the fact of her pain may likewise be thought to directly make a claim on us, where the validity of this claim in no way depends on our having validated it or on our having granted her the authority to make claims on us more broadly."[49]

Tarasenko-Struc does not—and I do not mean to—assume that if the ultimate ground of a duty is a fact about an agent (rather than another subject), then that duty cannot be genuinely directed at another subject. The heart of the argument is that if the explanatory ground of a theory of obligations is a self-relation, more must be said about how a self-relation can generate a self-standing self–other relation. Korsgaard's own strategy is not promising. It does not follow that the trick cannot be accomplished, but working through Tarasenko-Struc's argument can provide a vivid example of how things can go wrong with accounting for the sociality of morality—how metaethics can lend itself to a form of social alienation. It can at least bring into view the shape of the problem, and put some pressure on agent-centered theorists of normativity to say more about how the self–other gap can be bridged.

Importantly, the problem is generated by the Kantian constructivist theory of normativity: the explanation the Kantian provides for the truth aptness of normative facts entails that those facts have a certain structure. They are ultimately facts about how we stand in relation to ourselves, and not about how we stand with respect to others. Other forms of agent-centered metaethics run a

special prosecutor, at the risk of collapsing a limited privilege into an absolute immunity.

[49] Tarasenko-Struc, "Kantian Constructivism and the Authority of Others," 88.

similar risk, if not a greater one: if normative facts are ultimately explained in terms of agents' desires or other psychological states it is even more difficult to see how to recover the status of the other as the one who stands to be wronged.

One way to put the general worry is that the *reason* relation that forms the basis of normativity has argument places for the fact (or consideration) that is a reason, the agent for whom it is a reason, the action it is a reason to do, and perhaps the context, but not for the other, the one to whom a directed obligation is owed. The other may have a corresponding reason for a reactive attitude associated with being wronged, and thus the directedness of the reason would be at least partly accounted for as a psychological correspondence.[50] But to account for directedness in terms of merely corresponding reasons is to hold obligors and obligees at a normative distance from one another: the difference between having a reason to ϕ and owing it to someone in particular to ϕ is not that the other happens to have a specific attitude, but that one thereby stands to the other as witnesses to the same relational fact. The rights correlative to duties do not just happen to line up with them; they are inextricably linked. They are different perspectives on the same moral nexus between persons—indeed they are often claimed to be the very same fact expressed in two different ways.[51]

We might try to accommodate this feature of directed obligations by putting the duty or right in the "fact" argument place: [that *A* owes it to *B* to ϕ] is a reason for *A* to ϕ, and the very same fact is also a reason for *B* to (e.g.) resent *A* if *A* does not ϕ, and the same pair of reasons could be described in terms of the fact [that *B* has a claim right against *A* that *A* ϕ], which is after all the same fact. This will only push the problem back a step, however; *A*'s reason to ϕ and *B*'s reason to resent *A* if *A* does not ϕ will be constituted by a common fact (a fact about *A*'s duty, i.e., *B*'s right), but *A* will not be normatively related to *B* in virtue of having this reason, in which *B* only features as part of the content (like a movie features in my prudential reason to see it—more on this example below in section 3.5), rather than as a normative relatum. Metaethics must do more than generate the reasons associated with directedness if it is to fully vindicate the importance of recognizing another as standing to one in a relation of right.

[50] Darwall uses reactive attitudes and the standing to hold them to explain directedness, but it is not clear whether he is in fact *reducing* directedness to this correspondence. He claims that the concepts of authority, accountability, obligation, and the second person, as well as of attitudes like blame and the reasons or standing to hold them, come together in a circle. He is thus not reducing relational concepts like obligation to monadic or psychological concepts like attitudes, but using all of them to explicate the others (see *The Second-Person Standpoint* and "Bipolar Obligation").

[51] E.g., Gilbert, "Scanlon on Promissory Obligation"; and Wallace, *The Moral Nexus*.

For one person to owe a directed obligation to another is for them to recognize the other as the bearer of a claim against them, which is to recognize the other as recognizing them as owing a directed obligation, and so on. Contained within the self-consciousness that one stands in a juridical relation of this kind with another is at least the implicit recognition of the other as recognizing oneself. (Of course some bearers of rights and obligations are unaware, so it does not follow from one person's having a right against another that the other is similarly self-conscious, but the logic of directed obligations involves at least unrealized mutual recognition.) This is what mutual practical openness comes to in the realm of rights, and it is what metaethics needs to explain at the risk of leaving us socially alienated.

3.3. *Two-Level Accounts of Directedness*

One strategy available to Korsgaard or, for that matter, any other agent-centered metaethicist, would be to invoke the distinction between the content and justification of a norm like promise keeping. Thus the fact that *A* owes it to *B* to keep her promise can be explained by the role that promise plays in, for example, the integrity of *A*'s agency *so long as the promise itself is an entity partially constituted by B*. The content of a norm (that a promise is directed at *B*) and the justification of that norm (that you need to follow it to successfully constitute yourself as an agent) operate at different levels.[52] This kind of two-level theory, often associated with contractualism or rule utilitarianism, is usually criticized on the grounds that higher-level theories that generate rules without directedness built in get the extensions wrong, failing to reliably pick out correlative rights holders, or that in the particulars they fail to actually *explain* the correlativity of rights and duties altogether.[53]

But there are reasons to worry that in principle no such theory of directedness can succeed in vindicating it on the terms relevant to the problem of social alienation. Here the question is not about evaluating an action recommended by a practice (as in the original Rawls argument), but a relation of authority putatively established by it. But because authority is a higher-order moral concept, rules and practices are transparent when it comes to authority in a way they are not when it comes to reasons for action: it is one thing to say that a rule or practice can create reasons, and another altogether to say that a

[52] Cf. Rawls, "Two Concepts of Rules."

[53] E.g., Scanlon, *What We Owe to Each Other* (contractualism); Hooker, "Promises and Rule Consequentialism" (rule utilitarianism). Also see, e.g., Wenar, "Rights and What We Owe to Each Other"; Gilbert, "Scanlon on Promissory Obligation"; and Woods, "The Normative Force of Promising" (criticisms based on extensions and explanatory insufficiency, respectively).

rule or practice can establish basic, and not merely conventional, relations of authority and accountability.

Two-level accounts can create a fiction in which the rule has a structure the underlying normative theory lacks, but to see whether it is more than a fiction we need to look at whether the underlying normative theory can make sense of the structure. Suppose that the lawmakers of a legitimate political authority delegate enforcement power to a private party well positioned to track malfeasance—say, Google, with its immense surveillance apparatus. (And suppose—however implausible—that the legislators are right to do so, perhaps because it is an important issue and the state cannot deal with it alone, and the procedure through which Google will enforce the law does not violate any civil liberties.) When a Google auditor knocks on your door to ask you a few questions, you may be obligated to answer, even morally obligated, and within the fiction established by the law you may have to act as if you owe this duty to the Google auditor. But if you refuse, you may not wrong the auditor, or Google itself—you may wrong the state, or your fellow citizens, or perhaps no one at all. Figuring out the party to whom you truly owed the obligation (if any) requires going outside of the convention to see how the relevant authority (political, moral, legal) works and under what conditions (if any) it can be legitimately transferred to a third party.[54]

In the promissory case I discussed above, part of what it means to say that the promissory obligation is directed at *B* is to say that *B* is the bearer of not only the correlative claim right but the *power* of waiver. It may be that we take ourselves to be bound by rules that by convention stipulate some other person as the obligee, but that does not establish a genuine transfer of authority over our actions. In order to see whether on a given theory this is possible we need to "pierce the veil" of the convention and see whether the underlying account of authority is compatible with transferring it, or only with agreeing to act as if we have. That is precisely the move that Tarsenko-Struc targets under the guise of a transmission-of-authority principle, as I have just reviewed.

3.4. Social Alienation and Agent-Centered Metaethics

The agent-centered metaethical theories that we saw provide the resources to answer the challenge of normative alienation face special difficulties in accounting for the sociality of morality. These views explain moral facts starting with attitudes or capacities indexed to the individual, or from the first-person perspective. They thus come along with certain commitments about the kinds of facts moral theory can rely on: principally, facts about individual agents, or

[54] For a longer discussion of what is essentially the same point in a legal context, see Murphy, "Purely Formal Wrongs."

facts about oneself. Insofar as they aim to capture the sociality of morality, in the sense I have been discussing here, they are in the position of trying to reconstruct relational facts out of individual-agent facts, and it is not clear that this can be done. They may be able to recover the reasons associated with directed obligations, but if they do so by making such reasons out to be psychological facts about individual agents, or explained in terms of self-relations rather than social relations, that will not be enough.

The Kantian, for example, begins with facts about the nature of agency as such. Then, in attempting to derive substantive moral facts, she has to somehow generate facts of the right kind. That is, she has to generate facts suitable to bring others into view in the right way and explain the moral nexus that (for example) joins bearers of correlative rights and duties. While my discussion of social alienation in moral theory is in some important respects heterodox, under some description this is an aim that Korsgaard herself endorses. She holds that there is a role for sociality in the characteristic exercise of agency: reflecting on essentially public reasons, or responding to the call of another. Even on her own terms it is not clear that her conception of agency is up to the task of grounding the sociality that appears as a *deus ex machina* in lecture 4.2 of *Sources*. More broadly, the sense in which agency is social for Korsgaard is, so to speak, inside out. What it is to be an agent is essentially characterized by the potential to stand in recognitive relations with others, if there are any: the reflective relation that one stands in to oneself as an agent (the "second-person within," as she puts it elsewhere[55]) is generalizable. By her own lights, then, relations to others are not built into agency. What the above discussion of directedness suggests is that Korsgaard's theory is inadequate to vindicate the irreducible sociality of morality, and it is this structural feature of her theory that I suspect explains why. There is widespread skepticism regarding Kantian constructivism's ability to make good on its explanatory ambitions, and the gap between its agent-centered explanatory structure and the sociality of morality provides a compelling diagnosis. Explaining sociality in morality is a desideratum that at least some agent-centered approaches to metaethics recognize, and they are not set up to have a natural way of doing so.

3.5. *The Solution: Other-Centered Metaethics*

The demand to appreciate the significance of others as external, as particular, and as subjects themselves is realized in the demand to fully appreciate the directedness of certain moral requirements. There is an important sense in which at least some of the time what morality consists in is not recognizing

[55] See Korsgaard, "Autonomy and the Second Person Within."

oneself as having a reason or bound by a law but recognizing and responding to the other *qua* other.

In one sense the upshot of this discussion is somewhat trivial: moral facts are, at least some of the time, facts about particular others, and the relations we stand in to them. But what I have been trying to bring out is that this is not just a matter of the content of normative facts, but of their form. The other must show up in practical thought *in the right way*. Consider the reason I have to see a movie I am likely to enjoy. The movie shows up in an account of what I have reason to do. But when I reflect on the reason I have to respect the bodily autonomy of the person sitting next to me on the bus, they appear in my practical thought in a different way from the movie I am likely to enjoy, or they ought to if I am fully appreciating them as an individual reality.[56] Social alienation is thus a problem for metaethics insofar as it is in part concerned with how normativity works, about its structure, and further insofar as many theories of normativity seem committed to ruling out any way for us to play the right sort of role in the normative lives of one another. While constructivist, subjectivist, relativist, and other agent-centered approaches to metaethics can claim some success in addressing normative alienation, it is more traditional forms of realism that are better positioned to provide the resources for addressing social alienation.

Existing realist-metaethical theories may not be able to accommodate irreducible directedness without substantial revisions. As we saw above, a "reasons-first" realism of the kind associated with Parfit and Scanlon runs into the

[56] In something like the way that there is a formal difference between the way a *de re* thought relates to a referent and the way a *de dicto* thought relates to the same one, perhaps we should say that my thought of an other *qua* other relates me to her in a way that my thought of a movie *qua* potential source of pleasure does not. Some philosophers have sought to capture this distinction by insisting on the importance of second-personal thought in ethics (most famously probably Darwall), and though I quibble with the assimilation of this difference to one of grammatical person, I am inclined to endorse something like this line. For attempts to push the discussion of the second person in a direction similar to the one I am trying to go here, see Zylberman, "The Very Thought of You"; and Haase, "For Oneself and toward Another." The discussion of the second person that gets the closest to what I am after appears in Moran's characterization of the relationship between parties to successful communication:

> The relevant incorporation of another perspective on one's act and including that in one's own understanding of it is not the same thing as taking an "outside" perspective on what one is doing, something that each of the parties could do separately. The speaker does not imagine a third-person perspective on her act but rather a second-person one, that of her addressee; in adjusting her performance to this perspective she is not speaking so as to be overheard by an observer, but rather inhabiting the perspective of a *shared* participant in a practice, *the shared consciousness of what they are doing together*. (Moran, *The Exchange of Words*, 144, emphasis added)

difficulty that the reason relation lacks an argument place required to account for the other person that stands to one in the relation of duty and right, and thus risks theorizing away the relationality of directed obligations.[57] An emphasis on one species or another of normative facts—facts about fittingness or value or the good—leaves one similarly ill equipped to make out the fundamentality of the moral nexus that joins an agent and the other. Such theories deliver impersonal facts about the world that feature in specifying an agent's relation to possible actions, attitudes, or aims, but are not obviously relevant to an agent's relation to another.

But the basic realist strategy of taking whatever normative ethics delivers and promising to vindicate it by augmenting the ontological inventory (or, in more quietest flavors, but granting the legitimacy of a certain quasi-metaphysical discourse), is in principle perfectly consistent with taking directed obligations and the moral nexus they saturate as a primitive feature of reality (ontological or discursive). Whatever discourse of duties, rights, and sociality emerges the realist can simply affirm as a description of how things really are. If that means positing a new kind of metaphysical relation, so be it.[58]

That moral thought is at least sometimes thought of another, and that this difference is more than one of merely which singular terms appear in a reason-stating sentence suggests that a metaethics adequate to capture the sociality of morality will be somehow *other centered*. The explanation for how we come to have moral reasons will have to revolve around other creatures, how

[57] Nagel is an interesting case of a realist who comes close to explicitly setting for himself a goal like what I describe as avoiding social alienation—what he calls "practical solpsism"—but his focus is on recovering motivation and normative grip, rather than on explaining how his view can accommodate anything like directedness in particular or irreducible sociality in general (see *The Possibility of Altruism*). In other words, the challenge he sets for himself is to address normative alienation, so he offers little by way of directly accounting for social alienation. Given that his metaethics is reasons first and his primary route to avoiding practical solipsism is through publicity, rather than anything in the neighborhood of practical openness to the other, he is more in Korsgaard's position than the generic "realist" I am imagining here, who faces the opposite problem. (Perhaps this should not be surprising, as he, like Korsgaard, associates his view with Kant.)

[58] This suggestion is not meant to be dismissive. As with the analogous problem of perceptual alienation I allude to in note 20 above, where the direct realist answer is to simply insist that when we open our eyes in a well-lit room it is the objects in it that we see (i.e., to which we are perceptually related) without positing any mediating representations, I think it is in perfectly good order to insist that the self–other relations disclosed through practical openness to the other are just as real as anything else. The limitation of quietist realism is, as far as I am concerned, that the agent-centered approaches are right to worry about normative alienation; it is not metaphysical scruples that pull me in their direction, but a dissatisfaction with an unexplained connection between the other so disclosed and the self as open to them.

things are with them, and how they stand with respect to us. This is no real challenge for traditional realists, who can accommodate any constraint on what the normative facts must be like by saying of those facts, "yes, and they are simply true, no further explanation required." But as we will see in the conclusion, agent-centered approaches to metaethics struggle to meet the same standard, and thus to address the threat of social alienation.

4. Conclusion

Avoiding normative alienation urges making some concession toward agent-centered approaches to explaining normativity. But any explanation of what reasons an agent has that derives them from facts about her will risk having started in the wrong place to ever bring the other into view as an individual reality. To start with an individualistic account of the source of normativity and wind up with a full-throated vindication of normative facts as facts about concrete others appears to involve crossing a gap. Theories of normativity that define themselves by the task of accounting for the significance of the other-*qua*-other, however, risk having started in the wrong place to ever bring the resulting normativity close enough to the individual agent to avoid the threat of normative alienation.

The attempt to reckon with normative alienation pulls in the direction of agent-centered metaethics (typically though not exclusively irrealist, broadly construed), while the attempt to reckon with social alienation pulls in the direction of other-centered metaethics (typically nonnaturalist realism).[59] It is difficult for a theory of normativity to avoid both normative alienation *and* social alienation, but not impossible.

Supposing that a satisfyingly non-alienated theory of normativity must be in some sense agent centered *and* other centered, it will not do simply to impose the conjunction of the two constraints. There is at least a superficial tradeoff, in that, to take the metaphor a bit literally, the theory can have one center or the other, but not both. Working out how these constraints can coexist involves getting clearer on what it would mean to "center" the agent or the other in a theory of normativity—something that up until now I have expressed largely by example. What is the sense in which Humeans "center" the agent as a bearer of desires or values, or that Kantians "center" the agent as a bearer of the capacity for practical reason, in their explanation of how there can be normative facts?

[59] Strictly speaking it may be that the threat of social alienation is better understood as pulling in the direction of other-centered *ethics*, but that other-centered ethics is hard to square with agent-centered metaethics, and rather easy to ground in nonnaturalist realist metaethics.

It is tempting to reach for metaphysical notions like "grounding" and "fundamentality," but in this case I think their use obscures more than it reveals. Yes, desires are explanatorily fundamental for the Humean, and the capacity for practical reason grounds normativity for the Kantian. But nothing in this metaphysical gloss entails that normativity cannot have more than one partial ground or that more than one thing cannot be fundamental. Yet it remains unclear how one's own desires *and* the individual reality of another could be at once fundamental to the explanation of a given moral fact, other than by stipulation. What is needed is not the mere conjunction but a synthesis, a self–other relation wherein the other *qua* other is invoked in an understanding of what it is for the self to be a self.

These remarks are programmatic at best, but rather than attempting to develop them in any detail at this late stage of the argument I would like to close by considering a couple of positive proposals for how this could be done, coming, respectively, from either direction. First, from agency to sociality.

I have used Kantian constructivism as a stalking horse throughout this paper, largely because there is so much that it gets right. Korsgaard in particular begins with the insight (not original to her, but one that she centers in her own story) that even if we could make sense of the "queer" entities Mackie has long been taken to cast doubt upon, their mere existence would not be enough unless we had some explanation of how they could get a grip on us. Further, she takes on board more or less the social aims I have argued are necessary.

In my view, she does not have the explanatory resources to reach them. She begins with an individualistic conception of agency, one articulated in terms of an individual agent's capacities, capacities in turn understood through the form of law. Laws, on this picture, are universal generalities. In applying a law to oneself, one arrives at an instance: if we all ought to ϕ, then *I* ought to ϕ. Where is the other in this picture? The generality of a law hints at the logical possibility of another, but the law would still be a law if I were the only one around for it to bind. Korsgaard begins with this individualistic conception of agency and attempts to derive a picture of morality that has a deep social structure, in which we are responsive to the calls of others, who simply by speaking reshape the normative space in which we deliberate. This project is generally regarded as a failure.

The solution, it seems to me, or at least a solution, would be to build sociality into the story at the ground level: agency. Conveniently, for those of us who look to the history of philosophy to discern the movement of ideas (as Korsgaard clearly does), this suggestion has already been articulated by Kant's own successors in the tradition of German idealism: Fichte and Hegel. Both argue, in different ways, that self-consciousness—which marks the distinction between animal locomotion and rational action—depends on standing in relations of mutual recognition with other self-conscious creatures. Such a view is independently motivated, in ways I do not have the space to consider here, but for present purposes the appeal is that it has the potential to fund a constructivist theory of normativity that could both explain the grip reasons have on agents and the grip agents have on one another.[60]

What about the other direction? The way to address normative and social alienation beginning with other-centered realism and recovering the connection between normativity and individual agents, I want to suggest, is by taking a cue from Iris Murdoch. I argue elsewhere that we can read Murdoch as looking for a way of locating normativity in the world—in particular in historically conditioned social relations between concrete individuals—rather than in the attitudes or choices of the agent, while at the same time holding that getting oneself in a position to be responsive to it is itself an achievement of agency.[61]

Murdoch's is in some ways the paradigm of what I have called an other-centered metaethics, in that, as I noted above, for Murdoch the key element in morality is seeing others clearly, escaping fantasy and self-focus, and getting directly in touch with the individual reality of others. However, for Murdoch it is equally important to emphasize that the development of a distinctive practical standpoint on the world is something that we continuously and actively cultivate and revise, and is thus in an important sense the realization of individual agency.[62] That moral self-awareness is, for Murdoch, awareness of how one stands with respect to concrete other persons addresses social alienation, and that arriving at this form of self-awareness is something we struggle to do explains what the reality of others has to do with us, thereby addressing normative alienation.

Whether through the Hegelian strategy, the Murdochian strategy, some combination of the two, or some other approach altogether, metaethics has its work cut out for it in capturing the sociality of morality and its connection to individual agents.[63]

New York University School of Law
jsamuel@nyu.edu

[60] See Samuel, "Toward a Post-Kantian Constructivism"; and Peterson and Samuel, "The Right and the Wren."

[61] See Samuel, "Thin as a Needle, Quick as a Flash."

[62] Or so I argue as against what I take to be the more common reading of Murdoch as *contrasting* an ethics of clear vision with one of agency (Samuel, "Thin as a Needle, Quick as a Flash").

[63] Thanks to Sophie Cote, Sandy Diehl, Eleanor Gordon-Smith, Nathan Howard, Nick Laskowski, Kathryn Lindemann, Christa Peterson, Aaron Salomon, Keshav Singh, Michael

References

Bagnoli, Carla. "Kantian Constructivism and the Moral Problem." *Philosophia* 44, no. 4 (October 2016): 1229–46.

Bedke, Matthew. "A Dilemma for Non-naturalists: Irrationality or Immorality?" *Philosophical Studies* 177, no. 4 (April 2020): 1027–42.

Blanchard, Joshua. "Moral Realism and Philosophical Angst." In *Oxford Studies in Metaethics,* vol. 15, edited by Russ Shafer-Landau, 118–39. Oxford: Oxford University Press, 2020.

Boyd, Richard. "How to Be a Moral Realist." In *Essays on Moral Realism,* edited by Geoff Sayre-McCord, 181–228. Ithaca, NY: Cornell University Press, 1988.

Brandom, Robert. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press, 1994.

Brewer, Bill. *Perception and Reason*. Oxford: Oxford University Press, 1999.

Brink, David. "Externalist Moral Realism." *Southern Journal of Philosophy* 24, no. S1 (Spring 1986): 23–41.

Darwall, Stephen. "Bipolar Obligation." In *Oxford Studies in Metaethics,* vol. 7, edited by Russ Shafer-Landau, 333–58. Oxford: Oxford University Press, 2012.

———. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press, 2006.

Dasgupta, Shamik. "Normative Non-naturalism and the Problem of Authority." *Proceedings of the Aristotelian Society* 117, no. 3 (October 2017): 297–319.

Enoch, David. "Can There Be a Global, Interesting, Coherent Constructivism about Practical Reason?" *Philosophical Explorations* 12, no. 3 (2009): 319–39.

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

Finlay, Stephen. "The Obscurity of Internal Reasons." *Philosophers' Imprint* 9, no. 7 (July 2009): 1–22.

———. "Responding to Normativity." In *Oxford Studies in Metaethics,* vol. 2, edited by Russ Shafer-Landau, 220–39. Oxford: Clarendon Press, 2007.

Gibbard, Allen. "Morality as Consistency in Living: Korsgaard's Kantian Lectures." *Ethics* 110, no. 1 (October 1999): 140–64.

Gilbert, Margaret. "Scanlon on Promissory Obligation: The Problem of Promisees' Rights. *Journal of Philosophy* 101, no. 2 (February 2004): 83–109.

Haase, Matthias. "For Oneself and toward Another: The Puzzle about Recognition." *Philosophical Topics* 42, no. 1 (Spring 2014): 113–52.

Thompson, and Dan Webber for feedback on earlier drafts and helpful conversations along the way.

Harman, Gilbert. *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press, 1977.

Hayward, Max. "Immoral Realism." *Philosophical Studies* 176, no. 4 (July 2019): 897–914.

Hegel, G. W. F. *Phenomenology of Spirit*. Translated and edited by A.V. Miller. New York: Oxford University Press, 1977.

Honneth, Axel. "Foreword." In Jaeggi, *Alienation*, vii–x.

Hooker, Brad. "Promises and Rule Consequentialism." In *Promises and Agreements*, edited by Hanoch Sheinman, 235–52. New York: Oxford University Press, 2011.

Jaeggi, Rahel. *Alienation*. Translated by Frederick Neuhouser and Alan E. Smith. Edited by Frederick Neuhouser. New York: Columbia University Press, 2014.

Julius, A. J. "Suppose We Had Produced as Humans." Unpublished manuscript.

Kant, Immanuel. *Critique of Practical Reason*. Translated by Werner S. Pluhar. Indianapolis: Hackett Publishing Company, 2002.

———. *Groundwork of the Metaphysics of Morals*. Translated and edited by Mary Gregor. New York: Cambridge University Press, 1997.

Korsgaard, Christine. "Autonomy and the Second Person Within: A Commentary on Stephen Darwall's *The Second-Person Standpoint*." *Ethics* 118, no. 1 (October 2007): 8–23.

———. "Normativity, Necessity, and the Synthetic *a Priori*: A Response to Derek Parfit. Unpublished manuscript. http://www.people.fas.harvard.edu/korsgaar/Korsgaard.on.Parfit.pdf.

———. "Realism and Constructivism in Twentieth-Century Moral Philosophy." *Journal of Philosophical Research* 28, suppl. (2003): 99–122.

———. *Self-Constitution: Agency, Identity, and Integrity*. New York: Oxford University Press, 2009.

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

Lenman, James, and Yonatan Shemmer. "Introduction." In *Constructivism in Practical Philosophy*, edited by James Lenman and Yonatan Shemmer, 1–17. Oxford: Oxford University Press, 2012.

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

Manne, Kate. "Locating Morality: Moral Imperatives as Bodily Imperatives." In *Oxford Studies in Metaethics*, vol. 12, edited by Russ Shafer-Landau, 1–26. Oxford: Oxford University Press, 2017.

———. "On Being Social in Metaethics." In *Oxford Studies in Metaethics*, vol. 13, edited by Russ Shafer-Landau, 50–73. Oxford: Oxford University Press, 2008.

Martin, M. G. F. "On Being Alienated." In *Perceptual Experience,* edited by Tamar S. Gendler and John Hawthorne, 354–410. New York: Oxford University Press, 2006.

McDowell, John. "Criteria, Defeasibility, and Knowledge." *Proceedings of the British Academy* 68 (1982): 455–79.

———. *Mind and World*. Cambridge, MA: Harvard University Press, 1994.

Moore, G. E. *Principia Ethica*. New York: Cambridge University Press, 1903.

Moran, Richard. *The Exchange of Words: Speech, Testimony, and Intersubjectivity*. New York: Oxford University Press, 2018.

Murdoch, Iris. *The Sovereignty of Good*. New York: Routledge, 2001.

———. "The Sublime and the Good." In *Existentialists and Mystics,* 205–20. New York: Penguin, 1997.

Murphy, Liam. "Purely Formal Wrongs." In *Civil Wrongs and Justice in Private Law,* edited by Paul B. Miller and John Oberdiek, 19–39. New York: Oxford University Press, 2020.

Nagel, Thomas. *The Possibility of Altruism*. Princeton: Princeton University Press, 1970.

———. *The View from Nowhere*. Oxford: Oxford University Press, 1986.

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

Peterson, Christa, and Jack Samuel. "The Right and the Wren." In *Oxford Studies in Agency and Responsibility,* vol. 7, edited by David Shoemaker, 81–103. Oxford: Oxford University Press, 2021.

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

———. "Internalism for Externalists." *Philosophical Issues* 19, no. 1 (October 2009): 166–81.

———. "Moral Realism." *Philosophical Review* 95, no. 2 (April 1986): 163–207.

Rawls, John. "Kantian Constructivism in Moral Theory." *Journal of Philosophy* 77, no. 9 (September 1980): 515–72.

———. "Two Concepts of Rules." *Philosophical Review* 64, no. 1 (January 1955): 3–32.

Ross, W. D. *The Right and the Good*. Oxford: Oxford University Press, 1930.

Samuel, Jack. "An Individual Reality, Separate from Oneself: Alienation and Sociality in Moral Theory." *Inquiry* (forthcoming). Published online ahead of print, July 16, 2021. https://www-tandfonline-com.proxy.library.nyu.edu/doi/full/10.1080/0020174X.2021.1948445.

———. "Thin as a Needle, Quick as a Flash: On Murdoch on Agency and Moral Progress." *Review of Metaphysics* 75, no. 2 (December 2021): 345–73.

———. "Toward a Post-Kantian Constructivism." *Ergo* 9, no. 53 (2023). https://doi.org/10.3998/ergo.3116.

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

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

Schacht, Richard. *Alienation*. New York: Doubleday, 1970.

Schafer, Karl. "Hume on Practical Reason: Against the Normative Authority of Reason." In *The Oxford Handbook of Hume,* edited by Paul Russell, 356–79. Oxford: Oxford University Press, 2016.

———. "Realism and Constructivism in Kantian Metaethics: The Kantian Conception of Rationality and Rationalist Constructivism." *Philosophy Compass* 10, no. 10 (October 2015): 702–13.

———. "Realism and Constructivism in Kantian Metaethics: Realism and Constructivism in a Kantian Context." *Philosophy Compass* 10, no. 10 (October 2015): 690–701.

Scheffler, Samuel. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations underlying Rival Moral Conceptions*. Oxford: Clarendon Press, 1982.

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

Setiya, Kieran. "Other People." In *Rethinking the Value of Humanity,* edited by Sara Buss and Nandi Theunissen, 314–36. Oxford: Oxford University Press, 2022.

Shafer-Landau, Russ. *Moral Realism: A Defence*. Oxford: Oxford University Press, 2003.

Smith, Michael. "Search for the Source." *Philosophical Quarterly* 49, no. 196 (July 1999): 384–94.

Stocker, Michael. "The Schizophrenia of Modern Ethical Theories." *Journal of Philosophy* 73, no. 14 (August 1976): 453–66.

Street, Sharon. "Constructivism about Reasons." In *Oxford Studies in Metaethics,* vol. 3, edited by Russ Shafer-Landau, 207–45. Oxford: Oxford University Press, 2011.

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

Sturgeon, Nicholas. "Moral Explanation." In *Essays on Moral Realism,* edited by Geoff Sayre-McCord, 229–55. Ithaca, NY: Cornell University Press, 1988.

Tarasenko-Struc, Aleksy. "Kantian Constructivism and the Authority of Others." *European Journal of Philosophy* 28, no. 1 (March 2020): 77–92.

Thompson, Michael. "What Is It to Wrong Someone? A Puzzle about Justice." In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz,* edited by R. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith, 333–84. Oxford: Clarendon Press, 2004.

———. "You and I." Talk presented at the Aristotelian Society, London, May 2012.

Tiffany, Evan. "How Kantian Must Kantian Constructivists Be?" *Inquiry* 49, no. 6 (November 2006): 524–46.

Travis, Charles. "The Silence of the Senses." *Mind* 113, no. 449 (January 2004): 57–94.

Walden, Kenneth. "Laws of Nature, Laws of Freedom, and the Social Construction of Normativity." In *Oxford Studies in Metaethics*, vol. 7, edited by Russ Shafer-Landau, 37–79. Oxford: Oxford University Press, 2012.

———. "Morality, Agency, and Other People." *Ergo* 5, no. 3 (2018): 69–101.

———. "Mores and Morals: Metaethics and the Social." In *Routledge Handbook of Metaethics*, edited by Tristram McPherson and David Plunkett, 417–30. New York: Routledge, 2018.

———. "Reason and Respect." In *Oxford Studies in Metaethics*, vol. 15, edited by Russ Shafer-Landau, 1–23. Oxford: Oxford University Press, 2020.

Wallace, R. Jay. *The Moral Nexus*. Princeton: Princeton University Press, 2019.

Wenar, Leif. "Rights and What We Owe to Each Other." *Journal of Moral Philosophy* 10, no. 4 (2013): 375–99.

Williams, Bernard. "A Critique of Utilitarianism." In J. J. C. Smart and Bernard Williams. *Utilitarianism: For and Against*. New York: Cambridge University Press, 1973.

———. "Internal and External Reasons." In *Moral Luck*, 101–13.

———. *Moral Luck*. New York: Cambridge University Press, 1981.

———. "Moral Luck." In *Moral Luck*, 20–39.

———. "Persons, Character, and Morality." In *Moral Luck*, 1–19.

Woods, Jack. "The Normative Force of Promising." In *Oxford Studies in Normative Ethics*, vol. 6, edited by Mark Timmons, 77–101. Oxford: Oxford University Press, 2016.

Zhao, Michael. "Meaning, Moral Realism, and the Importance of Morality." *Philosophical Studies* 177, no. 3 (March 2020): 653–66.

Zylberman, Ariel. "The Very Thought of You." *Philosophical Topics* 42, no. 1 (Spring 2014): 153–75.

Journal of Ethics and Social Philosophy
vol. 26, no. 1 · november 2023

https://doi.org/10.26556/jesp.v26i1.2190

FAIRNESS AND CHANCE IN DIACHRONIC LOTTERIES

A RESPONSE TO VONG

Marie Kerguelen Feldblyum Le Blevennec

One influential view concerning the fairest way to distribute scarce or indivisible goods, especially between people in an equal or roughly equal situation, is that such goods should be distributed through a lottery.[1] In this paper, I focus on a debate about the way lotteries ought to be run in order to be fair. John Broome's synchronic account of lotteries has been criticized by Gerard Vong for being unfair in temporally extended cases. Vong holds that in order to be fair, even in such cases, lotteries must be diachronic rather than synchronic, and he offers his own account of how diachronic lotteries ought to be run. I will show that although Vong's diachronic account of lotteries is more plausible than Broome's original synchronic account, Vong's reply to a subsequent objection by Broome is implausible. This suggests that Vong's diachronic account must be modified in light of Broome's objection in order to treat all claimants fairly in temporally extended cases. I conclude by proposing one way to modify Vong's account to this effect.

Broome's account of lotteries for scarce or indivisible benefits focuses on giving claimants an equal chance to win a particular lottery at a single moment, which makes his account *synchronic*. Against Broome, Vong argues that synchronic lotteries are unfair because in temporally extended cases (during which new claimants can appear or the availability of goods can change), morally irrelevant factors can influence one's chances of winning a synchronic lottery. To illustrate this, Vong gives an example called Stormy Seas.[2] Two sailors, *A* and *B*, fall into the ocean during a storm, and only one can be saved because there is only one buoy on the ship. In order to determine in a fair way who will get the buoy, a lottery is run, and sailor *B* wins. Then, just before the buoy is thrown

[1] See, e.g., Broome, "Uncertainty and Fairness," 627–28, and "Fairness," 87; Burgers, "Perspective on the Fairness of Lotteries," 209–15; Sher, "What Makes a Lottery Fair?," 203.

[2] Vong, "Fairness, Benefiting by Lottery," 473.

to *B*, two other sailors, *C* and *D*, also fall into the water. Vong claims that on Broome's synchronic account of lotteries, another lottery would then have to be run between *B*, *C*, and *D*, with each having an equal chance to win.[3] According to Vong, Broome's account of lotteries is *procedurally* unfair in temporally extended cases like Stormy Seas because there are significant differences in the chance each claimant has to benefit, despite there being no morally relevant factors in the situation to justify this difference. In Stormy Seas, Broome's account would give *A* and *B* each a one-sixth chance of winning the buoy, whereas *C* and *D*, who fall into the water later and only participate in the second lottery, would each get a one-third chance of winning. As Vong correctly points out, this means that a major factor influencing one's chances of winning in the Broome version of Stormy Seas is *how late one falls into the water*, which is clearly morally irrelevant. So, although Broome's account provides all the sailors with an equal chance of winning *each* lottery, it does not provide them with an equal chance of winning the benefit precisely because his account is synchronic.[4]

To avoid this unfair allotment of chances, Vong suggests moving to a *diachronic* account of lotteries, which he calls the dual structure view.[5] He begins by distinguishing between *benefit* and *procedural* claims.[6] When someone loses

[3] On Broome's account, *A* is not included in the second lottery because *A* has lost his claim to the buoy by participating in the first lottery (which he loses against *B*) and getting surrogate satisfaction for his claim. A claim is surrogately satisfied when the individual holding it had the chance to benefit from the good by participating in a fair distribution procedure for that good; see Broome, "Fairness," 98.

[4] Here, it is assumed that what makes a procedure fair is that it provides an equal chance of *benefiting from the good* to all participants, not merely that it gives an equal chance of winning a particular lottery. Winning a particular lottery is not what matters here; what is at stake is the *distribution* of the specific good. This is why, for Vong, temporally extended cases such as Stormy Seas show that "Broome's synchronic view of claim satisfaction by lottery undermines his view of fairness [according to which each individual should have an equal chance of benefiting] in temporally extended cases such as Stormy Seas" (Vong, "Lottery," 477).

[5] See Vong, "Lottery," 471, 479. There are numerous possible diachronic accounts of lotteries (for example, "diachronic weakening" or "diachronic strengthening"—see Vong, "Lottery," 478). However, here I will focus on Vong's dual structure view, which seems to me to be more plausible than the other diachronic accounts he mentions. Explaining why in detail is beyond the scope of this paper. For more details about why the other diachronic accounts Vong mentions are not as plausible as Vong's dual structure view, see his discussion of the other diachronic accounts ("Lottery," 477–79).

[6] Procedural claims are one's claims "not to be treated inappropriately, … which is a matter of procedural *ex ante* fairness. … It is these *procedural* claims that are satisfied by a lottery that gives claimants appropriate, fair chances of benefiting" (Vong, "Lottery," 479). In contrast, *benefit* claims "are satisfied, *ex post*, when the benefits are actually distributed" and "cannot be lost due to the results of a procedurally fair lottery" (479–80).

a particular lottery, he may have his procedural claim satisfied, but he does not lose his *benefit* claim and so should be included in any new lottery where that benefit can be won. Next, Vong claims that once *C* and *D* fall into the water after the first lottery between *A* and *B*, the results of that first lottery should be ignored. This is because the procedure according to which it was run can no longer be considered fair once *C* and *D* are in the water—there are now new claimants, rendering the first lottery no longer procedurally fair.[7] According to Vong, *A* should be included in the second lottery between *B*, *C*, and *D*, as *A* did not lose his *benefit claim*. Moreover, since all the sailors have a benefit claim, each should have an equal chance of winning this second lottery so that their procedural claims are respected. This would ensure that each sailor has an equal *overall* chance to benefit from the good and that the diachronic lottery is fair.[8]

Table 1. Stormy Seas

	Chances to Win	
Lottery	Vong's Account	Broome's Account
Lottery 1 (t_1)	A = 50% B = 50%	A = 50% B = 50%
Lottery 1 Result	B wins	B wins
Lottery 2 (t_2)	A = 25% at t_2 (¼ chance overall) B = 25% at t_2 (¼ chance overall) C = 25% at t_2 (¼ chance overall) D = 25% at t_2 (¼ chance overall)	A = 0% at t_2 (⅙ chance overall) B = 33% at t_2 (⅙ chance overall) C = 33% at t_2 (⅓ chance overall) D = 33% at t_2 (⅓ chance overall)

Note: "Overall" = across t_1 and t_2.

At first glance, Vong's diachronic dual-structure view of lotteries seems fairer than Broome's synchronic account, as on Vong's view, each sailor has an equal *overall* chance of getting the buoy. However, Broome has voiced a worry about Vong's account. For Broome, the fact that *B* does not have his win of Lottery 1 taken into account in Lottery 2 is problematic.[9] It seems *B* could *justifiably* complain that it is unfair that his victory in Lottery 1 is not recognized, especially if *A* is included in Lottery 2 despite losing in Lottery 1 (as Vong claims *A* should be) and then ends up winning Lottery 2.[10] The upshot is that Vong's diachronic

[7] See Vong, "Lottery," 483.

[8] Note that, unlike Vong, Broome does not distinguish between benefit claims and procedural claims when he uses the term "claim."

[9] See Vong, "Lottery," 481.

[10] See Vong, "Lottery," 481.

account of lotteries seems unfair because it does not give *any* recognition to winners of lotteries prior to the final lottery.

To see why Broome finds this unintuitive, consider the following case. Sailors *A* and *B* fall overboard, and a lottery is run to decide who gets the only buoy on board the ship. *B* wins the first lottery against *A*, but then before *B* gets the buoy, *C* and *D* fall into the water. A new lottery is run according to Vong's view so that *A* is included despite losing the first lottery, and *A* wins the second lottery. It seems natural that *B* would be upset that *A* gets the buoy instead of himself. In fact, we can imagine a scenario in which *B* wins every lottery, but after each win, new claimants join the lottery process (i.e., more sailors fall into the water), so a new lottery is run. *B* then loses the final lottery. In such a scenario, *B* could have won several lotteries and lost only one, but because the lottery he loses happens to be the one after which no new claimants appear, he does not get the good (and potentially dies).[11]

Vong's response to this worry is to deny that in diachronic lotteries, a winner of a previous round has any special claim in a subsequent round. To show why, Vong provides an example: the Defective Extra Buoy case (DEB).[12] In DEB, two sailors, *A* and *B*, fall into the water, and the only buoy available is defective—it can save one of the sailors but will cause him considerably more stress in the process than a normal buoy would. A lottery is run to determine who will get this defective buoy, which *B* wins. However, before the buoy is thrown to *B*, the captain discovers an extra, non-defective, normal buoy on board the ship. The captain throws the regular buoy to *B* and the defective buoy to *A*.

Vong argues that for *B* to justifiably complain in such a scenario, his complaint has to concern his not receiving *specifically* what he won in the earlier lottery—namely, the defective buoy; *B*'s complaint would be that his claim to a *specific benefit* was not respected. Vong calls the notion that *B* has "a claim not just on the general benefit of having their life saved, but on the specific benefit of having their life saved by the first buoy" the "specific benefits view."[13] Vong rejects the specific benefits view because it seems intuitively absurd for *B* to complain about not getting the specific buoy he won (the defective buoy). Vong's intuition is that *B*'s benefit claim is satisfied simply by being saved, not

[11] The same applies to an extended version of Stormy Seas, in which a large ship gets caught in a storm; two sailors are washed overboard, but every time the captain runs a lottery, more sailors fall into the water before he distributes the benefit. On Vong's view, a sailor who wins every lottery except one could end up dying simply because the lottery he loses happens to be the one after which no new claimants appear.

[12] See Vong, "Lottery," 481–82.

[13] Vong, "Lottery," 474–75.

by getting a specific buoy.[14] Since it would be absurd for *B* to complain, Vong concludes that *B* actually does not have a legitimate complaint if his win of a previous lottery is not recognized in subsequent lotteries.

However, it seems to me that Vong's DEB example does not convincingly support his argument that the winner of a fair lottery lacks grounds for a legitimate complaint if his previous win is not taken into account in a subsequent lottery. This is because DEB ignores a crucial difference between two types of reasons to complain.

When someone complains, we can distinguish two types of reasons for their complaint: (i) *claim-based* reasons to complain and (ii) *normative* reasons to complain in general. Claim-based reasons to complain are one kind of legitimate reason to complain, which are directly relevant to an agent's claim on a particular benefit in virtue of which the agent's complaint would be fitting. In contrast, normative reasons to complain in general are simply those reasons on the basis of which it is worth complaining at a given moment, regardless of whether the complaint is legitimate and fitting in terms of a genuine claim that one is owed something. One example of a normative reason to complain that is not a claim-based reason to complain is that complaining in a given situation would be in one's self-interest (regardless of what claims one happens to have in the situation). For example, in a case where only *A* and *B* are drowning, and *B* wins a fair lottery, *A* might still have other *normative* reasons to complain out of self-interest in order to try to convince the ship captain to throw him the buoy so that he can survive, even if *A* has no *claim-based* reasons to complain because *A* lost a fair lottery.

The important point here is that claim-based reasons to complain and other normative reasons to complain sometimes come apart. That is, claim-based reasons to complain are *one kind* of normative reason to complain, which means that one can be rationally motivated to complain even when one lacks a claim-based reason to complain. Moreover, claim-based reasons to complain can be outweighed by other normative reasons to complain. This means that one can be rationally motivated *not* to complain even when one has a claim-based reason to complain because one's claim-based reason is outweighed by one's other normative reasons regarding the option of complaining.[15]

[14] See Vong, "Lottery," 475.

[15] Of course, sometimes people are motivated to complain by factors that are not normative reasons at all—for instance, if someone is delusional or suffering from *akrasia*. However, when I talk about normative *reasons* to complain, I am talking about factors that it would intuitively be reasonable to take as genuine reasons worth complaining on the basis of. My goal here is just to distinguish the narrow category of claim-based reasons to complain from this broader category of normative reasons to complain.

Once we apply this distinction between claim-based and normative reasons to complain to DEB, Vong's response to Broome's objection becomes less convincing. In DEB, the reason *B* does not complain is not necessarily because he lacks claim-based reasons to do so, as Vong contends. In fact, it seems more plausible to say that *B* does not complain simply because he knows that he will be better off with the normal buoy than with the defective one. *B*'s normative reason not to complain outweighs his claim-based reason to complain, so he does not complain.

To see that this explanation is more plausible, imagine a reversed scenario in which *B* wins a regular buoy in the lottery but is instead given an extra, defective buoy. Intuitively, in such a scenario, it would not seem absurd for *B* to complain. Rather, it seems plausible that once *B* is back on board, stressed and gasping for breath after having barely been saved by the defective buoy while *A* stands calmly after having been smoothly rescued with the regular buoy, *B* might justifiably complain about having received the defective buoy instead of the regular buoy he won, despite having had his life saved. It is not absurd to imagine *B* asking: "Why didn't you throw me the buoy I won? Why did you give it to *A* instead?" That is, in this reversed scenario, it does not seem absurd for *B* to appeal to a claims-based reason to justify his complaint about not receiving the specific benefit he won, despite having received the general benefit of having his life saved.

This intuition seems easy to explain if we accept Broome's suggestion, on which *B* has a *specific* claim on the regular buoy due to his lottery win.[16] In contrast, for Vong to be able to explain this intuition about the reversed case while maintaining that *B* lacks a claims-based reason to complain in DEB, it would have to be the case that whether the buoy *B* complains about is regular or defective is a morally relevant factor, since that is the only difference between DEB and the reversed case that could explain the difference in the legitimacy and fittingness of *B*'s complaint in the two scenarios. Yet this move is not available to Vong if, as he contends, *B*'s claim is just on the general benefit of having his life saved; whether the buoy is regular or defective in the manner described in DEB does not affect the benefit Vong thinks is at stake.

Since Vong's position faces this difficulty, it seems more plausible to say as I do that in both cases, *B* does have a claim-based reason to complain, but that in DEB, his claim-based reason to complain and his other normative reasons regarding the option of complaining come apart, while in my reversed case they are in harmony. In other words, in DEB, *B* clearly has other normative reasons *not to complain*, and these other normative reasons happen to outweigh the

[16] Vong, "Lottery," 474–75.

claim-based reasons he possesses to complain about not getting what he specifically won in the lottery. So, DEB does not rule out the possibility that *B* has claim-based reasons to complain. As a result, we can claim against Vong that *B* would be, in an important sense, *justified* in complaining about not receiving the specific (defective) buoy that he won, even if we also acknowledge that it would be pragmatically absurd for *B* to *actually* do so.

Vong has offered the following reply to my argument in correspondence: "To better understand the effect of legitimate reasons alone, we need a case ... which ... shows that when there are no instrumental reasons at play, there still is not any justified reason for complaint on behalf of someone who 'won' a different specific good than the one that they ultimately received that was equally as good." The scenario Vong suggests here corresponds to one of the examples in his article: the Extra Buoy case. This case is similar to DEB, except for the fact that both buoys are regular buoys—we can imagine that they differ merely in color, with the first buoy being red and the extra buoy being blue. For Vong, it would be absurd to insist on giving the red buoy to the lottery winner *B* and the blue buoy discovered post-lottery to the lottery loser *A*: "Intuitively, it does not matter who gets which buoy, as long as both of their lives are saved."[17] However, in cases like this, I am willing to simply bite the bullet and claim that, *strictly speaking*, there are reasons to give specifically the red buoy to *B* and the second blue buoy to *A* as opposed to the other way around, even if in practice there is no need to criticize the captain if he fails to do this or to compensate *B* if this is not done. The fact is that sailor *B* participated in a procedure that is *designed* to provide the winner with a good that has been specified in advance—in this case, the red buoy. This fact does not change simply because, in a case like Extra Buoy, *B* lacks any pragmatic reasons to complain about getting the otherwise identical blue buoy.

In sum, Vong's reply to Broome, which rejects the specific benefits view, yields plausible answers in Vong's original DEB case and the Extra Buoy case (including when the two buoys are different colors). But it has trouble with my reversed DEB case because, given that Vong thinks the benefit in question is the general one of having one's life saved, Vong would have trouble explaining why, intuitively, one has a claim-based reason to complain if one gets a defective buoy instead of the specific regular buoy one wins in the lottery. On the other hand, my proposal, which accepts the specific benefits view and distinguishes between different reasons to complain, yields intuitively plausible answers in both my reversed DEB case and Vong's original DEB case. But it requires biting the bullet in the Extra Buoy case (including when the two buoys are different

[17] Vong, "Lottery," 475.

colors) and claiming that, strictly speaking, one has a claim-based reason to complain even in the Extra Buoy case. It seems to me that Vong lacking a plausible explanation for why one has a claim-based reason to complain if one gets a defective buoy rather than the specific normal buoy one wins in the lottery is a bigger problem than the bullet I have to bite in the Extra Buoy case (including when the two buoys are different colors). It is more important to successfully account for all instances of legitimate claim-based complaints than it is to avoid trivial claim-based complaints.

In light of all this, there is reason, after all, to remain unconvinced by Vong's response to Broome's objection. Intuitively, a winner of a previous lottery round does have a special claim in a subsequent round and can justifiably complain if this is not taken into account.

Although Vong's dual structure view of lotteries is problematic in light of Broome's objection, I do not think it should be fully rejected, especially since Vong's basic argument that lotteries should be diachronic rather than synchronic seems right. Instead, I want to modify Vong's model in order to account for Broome's intuitively plausible suggestion that, for diachronic lotteries to be fair, subsequent lotteries should take "into account earlier results."[18] One way to integrate Broome's intuition into Vong's account in a principled way could be to include in subsequent lotteries all individuals who have a benefit claim *except* those who have already had a procedural claim satisfied *with respect to that specific benefit* by losing a previous lottery. That way, we can mostly capture Vong's intuition that if you have a benefit claim, you should be included in a lottery for that benefit, while also capturing Broome's intuition that having a procedural claim satisfied can impact your benefit claim.[19]

Moreover, Broome has suggested to Vong a way to run "subsequent lotteries between previous winners and *subsequent claimants*, while adjusting the probabilities of subsequent lotteries to ensure all claimants have an appropriate chance of benefiting."[20] This suggestion is useful for implementing the modifications to Vong's account that I have proposed. For example, in the Stormy Seas case, Broome proposes running a *weighted* diachronic lottery in which *B* is given an increased chance of winning Lottery 2, such that his winning Lottery 1 is recognized *and* such that *B*'s *overall* chances of winning the benefit are equal to those of all other claimants in Lottery 2.[21]

[18] Vong, "Lottery," 481.

[19] This intuition is reflected in Broome's notion of surrogate satisfaction of claims; see Broome, "Fairness," 98.

[20] Vong, "Lottery," 481.

[21] See Vong, "Lottery," 481.

Table 2. Stormy Seas Revisited

	Chances to Win	
Lottery	Vong's Account	Weighted Diachronic Account
Lottery 1 (t_1)	A = 50% B = 50%	A = 50% B = 50%
Lottery 1 Result	B wins	B wins
Lottery 2 (t_2)	A = 25% at t_2 (¼ chance overall) B = 25% at t_2 (¼ chance overall) C = 25% at t_2 (¼ chance overall) D = 25% at t_2 (¼ chance overall)	A = 0% at t_2 (¼ chance overall) B = 50% at t_2 (¼ chance overall) C = 25% at t_2 (¼ chance overall) D = 25% at t_2 (¼ chance overall)

Note: "Overall" = across t_1 and t_2.

In the kind of weighted diachronic lottery proposed here, in each new round, the winner of the previous round is given the chances that the losers of the previous round would have had in the new round. So, in the Stormy Seas case, if sailor *E* happens to fall in the water after Lottery 2 (which *B* wins) and before the buoy is given to *B*, a third lottery must be run between *B* and *E*, in which *B* is given the chances that *A*, *C*, and *D* would have had in the new lottery (if everybody had an equal chance). This would result in *B* having a four-fifths chance to win Lottery 3, whereas *E* would have a one-fifth chance to win Lottery 3. However, *B* and *E* would each have an *overall* chance of one-fifth to win the buoy (across lotteries 1–3).

In his article, Vong rejects Broome's new suggestion of a weighted diachronic lottery due to his argument based on DEB. However, since I have shown why that argument is unconvincing, I contend that the account I have given here based on Broome's suggestion is the fairest option as it avoids the two problems that we have discussed: (i) morally irrelevant features such as how late one falls into the water, have no influence on a claimant's chances of winning, and (ii) *B*'s winning Lottery 1 (and *A*'s losing Lottery 1) is not ignored.[22]

At first glance, one might think that a certain variant of Vong's Stormy Seas case could pose a problem for this account of weighted diachronic lotteries. Imagine a case in which, at first, only one sailor, *A*, falls into the water, and the captain runs a lottery just for *A*, which *A* has a 100 percent chance of winning and which *A* wins. However, just before *A* is given the buoy, sailors *B*, *C*, and

[22] Some might argue that *A* could complain about not being included in the second lottery. However, I contend that although *A* might have some normative reason to complain, in general he does not have any *claim-based* reasons to do so (or any other *legitimate* reason to complain). After all, *A*'s overall chance of winning the benefit across the two lotteries is the same as that of each other sailor. Moreover, *A*'s procedural claim is satisfied through his participation in a fair distribution process.

D all fall into the water, and a new lottery is run. It seems like a supporter of my view would have to claim that *A*'s win of the first lottery must be ignored in order for *A*, *B*, *C*, and *D* all to have an equal overall chance (across both lotteries) at receiving the buoy. Yet ignoring *A*'s win in the first lottery is precisely the outcome my view was supposed to avoid.

However, I do not think this example poses a genuine problem for my view. The point of running a lottery in the first place is to fairly distribute a scarce or indivisible good among multiple claimants. When there is only one claimant for one good, or when the goods are abundant enough or divisible such that all claimants can be satisfied, there is simply *no need* for a lottery, whether morally or practically. So, in a case like the one we are considering here, the first lottery run for a single claimant has no moral or practical relevance and, therefore, need not be taken into account when the second lottery is run for multiple claimants. In other words, there is no morally relevant factor from the first lottery that needs to be taken into account in the second lottery precisely because the first lottery was frivolous. Thus, there is no problem with assigning an equal 25 percent chance to all of *A*, *B*, *C*, and *D* in the second lottery because there is not actually anything from the first lottery that is relevant to the question of the second lottery's fairness.[23]

Another potentially problematic case is one in which there *is* moral and practical reason to run a lottery in the first instance because there are multiple beneficiaries, but one of the potential beneficiaries becomes no longer available to receive the benefit. For example, consider a case in which *A* and *B* fall into the water, a lottery is held, but *A* drowns. This sort of case can take one of two forms, depending on when *A* becomes no longer available to receive the benefit—that is, depending on whether *A* dies *after the first lottery has concluded* and *A* has been declared the winner or dies *prior to this while the first lottery is still going on*.

If *A* dies while the first lottery is still going on, that lottery would cease to reflect the actually existing claims on the benefit. Since *A* is dead, *A* no longer has a claim, which means that only *B* (who is still alive in the water) has a claim on the benefit, yet the lottery is structured as though there are two claims on the benefit. This means that in the scenario where *A* dies *during* the first lottery, the first lottery must be nullified. At that point, as long as no more sailors have

[23] More generally, it is also worth emphasizing that given the practical and moral question that lotteries are meant to solve, it simply does not make sense to run a lottery when there is only one claimant. To do so is to implement a solution without a problem. Moreover, in certain cases where the stakes are high, such as life or death cases like the variant of Stormy Seas we are considering, it might even be *immoral* to run a lottery when there is only one claimant and he is drowning. It seems like what we ought to do morally is to give the lone claimant the buoy as quickly as possible.

fallen overboard yet, then on my view, the resulting scenario is exactly like the one discussed above where only *B* is overboard—there is no need to hold a second lottery. On the other hand, if more sailors fall into the water at this point, a new lottery should be held, giving equal chances to *B* and these other sailors. Either way, as discussed above, there would be no problem for my view.

Alternatively, if *A* dies after the first lottery has concluded and *B* has been declared the winner, then on my view, once more sailors fall into the water, the scenario is exactly like the normal Stormy Seas case: a second lottery is held and *B* is given the chances that A *would have had* in the second lottery. Recall that, on my view, in each new round the winner of the previous round gets the chances that the losers of the previous round would have had in the new round. The fact that *A* actually happens to have died after the first lottery has concluded does not change how the second lottery should be run—the second lottery should still take into account that counterfactual about the chances *A* would have had in the second lottery. This is because the reason *B* is supposed to get the chances the loser of the previous lottery would have had is to reflect *B*'s win of the first lottery while maintaining the same overall chance of winning the benefit for all the sailors who fell overboard. Whether *A* dies after losing the first lottery or survives does not affect this consideration. So, whether *A* dies during or immediately after the first lottery, the situation will be identical to one of the cases that my view can deal with without a problem—either the case discussed above where only *B* falls in the water, or the original Stormy Seas case.

In sum, my proposed account of lotteries combines the advantages of both Vong's and Broome's accounts while avoiding their disadvantages. Unlike Broome's original account, my proposed account is diachronic and thus avoids appealing to morally irrelevant factors while providing equal overall chances of winning to anyone with a claim on a given benefit in temporally extended cases (which makes it procedurally fair according to Vong's standards). And unlike Vong's account, my proposed account recognizes that having procedural claims satisfied can change one's benefit claims in subsequent lottery rounds (as in my proposed treatment of *A* in Stormy Seas) and that winning previous lottery rounds should be taken into account in temporally extended cases (as in my proposed treatment of *B* in Stormy Seas). That is, my new account is diachronic and genuinely procedurally fair, which makes it more plausible than Broome's synchronic account or Vong's dual structure view. It takes into account the way changes in the number of claimants or availability of goods in temporally extended cases can affect the fairness of a lottery run prior to those changes, but without totally ignoring the results of those prior lotteries. Hopefully, all this can contribute to finding the fairest way to run lotteries, which is particularly important as it can have a significant impact on people's lives—especially in cases related to health where there are often not enough goods to satisfy everybody, such as organ distribution or Medicaid lotteries.[24]

Boston University
marieklb@bu.edu

REFERENCES

Allen, Heidi, Katherine Baicker, Sarah Taubman, Bill Wright, and Amy Finkelstein. "The Oregon Health Insurance Experiment: When Limited Policy Resources Provide Research Opportunities." *Journal of Health Politics, Policy, and Law* 38, no. 6 (December 2013): 1183–92.

Broome, John. "Fairness." *Proceedings of the Aristotelian Society, New Series,* 91 (1990): 87-101.

———. "Uncertainty and Fairness." *The Economic Journal* 94, no. 375 (September 1984): 624–32.

Burgers, Jan-Willem. "Perspectives on the Fairness of Lotteries." *Res Publica* 22, no. 2 (May 2016): 209–24.

Sher, George. "What Makes a Lottery Fair?" *Noûs* 14, no. 2 (May 1980): 203–16.

Vong, Gerard. "Fairness, Benefiting by Lottery and the Chancy Satisfaction of Moral Claims." *Utilitas* 27, no. 4 (December 2015): 470–86.

[24] I am grateful to Gerard Vong for bringing my attention to cases such as the Oregon Medicaid lottery. See also Allen, Baicker, Taubman, Wright, and Finkelstein, "The Oregon Health Insurance Experiment," 1183–92.

Journal of Ethics and Social Philosophy
VOL. 26, NO. 1 · OCTOBER 2023

https://doi.org/10.26556/jesp.v26i1.2915

THE CASE FOR VOTING TO CHANGE THE OUTCOMES IS WEAKER THAN IT MAY SEEM

A REPLY TO ZACH BARNETT

Amir Liron and David Enoch

You are unlikely—*really* unlikely—to cast a deciding vote in the next election. Why vote at all, then? You may have reasons to vote that are not sensitive to how likely you are to change the outcomes, but let us put them to one side for now.[1] Do you have a reason to vote to change the outcomes? Whether you do depends, of course, not only on the probability of affecting the outcomes but also on the payoff if you do. For it to be rational to vote (to change the outcomes), it has to be the case that the expected value of voting (roughly, your chances of changing the outcomes multiplied by the significance of the change you will make) is higher than the cost of voting. It is a very common view that this is hardly ever the case and that you are almost always in a position to be all but certain that this will not be the case.[2] Disappointingly, this is said to be so even if you care about the common good so that the size of the payoff in the above inequality, if you do end up making a difference, includes the good outcomes for all—not just for you.

In a recent paper, Zach Barnett forcefully argues that this is a mistake. He shows how it follows, from rather conservative assumptions, that in many real-life cases, the expected social value of voting is higher than its (personal, and so presumably also social) cost, so that at least for a voter who is motivated to promote the common good, it does make sense to vote in order to change the outcomes.[3] Barnett is successful, we believe, in showing that the commonly held belief (that voters, because so unbelievably unlikely to make a difference,

[1] For an initial discussion, and for references, see Brennan, "The Ethics and Rationality of Voting," sec. 1.3.

[2] See Brennan, "The Ethics and Rationality of Voting," sec. 1.1., and the many references there. See also the quotes from Brennan in Barnett, "Why You Should Vote to Change the Outcome."

[3] Barnett, "Why You Should Vote to Change the Outcome."

do not have a reason to vote in order to change the outcomes) is way too hasty. And this, despite our criticism below, is, of course, a significant achievement.

However, Barnett is—we argue here—too quick on one key premise, and once this is noticed, it is not clear how often Barnett's reasoning can point to a justification of voting to change the outcomes. Indeed, the problem facing Barnett here is very similar to what is arguably the underlying problem with the more pessimistic models he rejects. In this way, Barnett's reasoning may apply to significantly fewer real-life scenarios than he suggests.

1. Barnett's Argument

It is important for Barnett (for reasons we return to below) to avoid relying on too theory-driven modeling assumptions here. Instead, his argument relies on rather specific and arguably plausible conditions:

> *Stakes Condition*: The average social benefit (b) per citizen of electing the better candidate is more than twice as great as the cost (c) of voting (in short: $b > 2 \times c$).
>
> *Chances Condition*: The probability of casting the deciding vote (d) is at least one divided by the number (N) of citizens (in short: $d \geq 1/N$).[4]

Given these two conditions, it trivially follows that the expected value of voting is higher than its cost (in short: $\frac{1}{2} \times b \times d > c$).[5] We will not question this derivation, of course. Nor will we question the Stakes condition: it probably does not hold in full generality (nor does Barnett argue that it does), but it does seem to hold for many people, in many elections, in at least reasonably well-run democracies, and we are happy to constrain the discussion to just those.[6] The question for us, then, is why should we accept the Chances condition?

Barnett relies on the following two premises:

> *Partial Unimodality*: The leading candidate is at least as likely to earn exactly half of the vote as she is to earn any precise share of the vote smaller than this.

[4] Barnett, "Why You Should Vote to Change the Outcome," 427.

[5] Unless, that is, one is uncertain regarding the right vote. The less certain one is that the difference one's vote will make (if indeed it makes a difference) is for the better, the less confident one should be that the expected value of one's vote is higher than its cost (for it may be negative). For a critique of optimistic suggestions about voting to change the outcomes (including Barnett's) that emphasizes this point, see Brennan and Freiman, "Why Swing-State Voting Is Not Effective Altruism." We put this kind of consideration to one side here.

[6] Barnett's relevant section is titled "The Stakes Condition: A Qualified Defense."

> *Narrow Upsets*: If the leading candidate fails to earn a majority, then the likelihood that she comes within ten percentage points of her opponent is at least ½.[7]

With these two assumptions in place, Barnett argues that as long as both candidates have at least a 10 percent chance of winning (surely, an easily met condition in real-world elections), the Chances condition follows—namely, that the probability of your vote being the deciding vote is greater than (or equal to) one divided by the number of voters ($d \geq 1/N$).

We will not take issue with this derivation. Nor will we doubt the Narrow Upsets premise, which—while, of course, is not a necessary truth—seems empirically very plausible, at least for the vast majority of elections.[8] The problem lies, we proceed to argue, with Partial Unimodality.

2. How Partial Unimodality May Fail

Partial Unimodality is, as Barnett notes, intuitively plausible. Suppose Daisy is projected to receive 52 percent of the votes. If so, then an outcome of 50 percent for Daisy is closer to the projection than any outcome where she receives fewer votes. Assuming that the likeliest outcomes are clustered together and that outcomes become less and less likely as one moves further away from the likeliest cluster, Partial Unimodality follows. This is especially clear if we assume (for now) normal distribution around the projected result.[9] As can be seen from figure 1, the further left from the 50 percent line, the lower the probability, so that Partial Unimodality is guaranteed to be true. And Barnett does not need something as strong as normal distribution. As long as the distribution of likelihood of results is sufficiently similar to the one in figure 1, Partial Unimodality is guaranteed to be true.[10]

[7] Barnett, "Why You Should Vote to Change the Outcome," 434.

[8] Of course, Barnett's reasoning applies—at least as-is—only to voting systems of the kind he describes. Whether the reasoning can be extended to other voting systems is an open question, to be answered piecemeal. We do not challenge Barnett on this front (and we thank an anonymous referee for relevant discussion here).

[9] Barnett does not assume anything as strong as normal distribution. In fact, he rejects this (when criticizing Brennan and the binomial model). We return to this. We rely here on the case of normal distribution for heuristic purposes alone.

[10] And there may be some case of a distribution that satisfies Partial Unimodality even though it is quite far from a normal distribution—like, for instance, uniform distribution. But we can safely ignore such cases here.

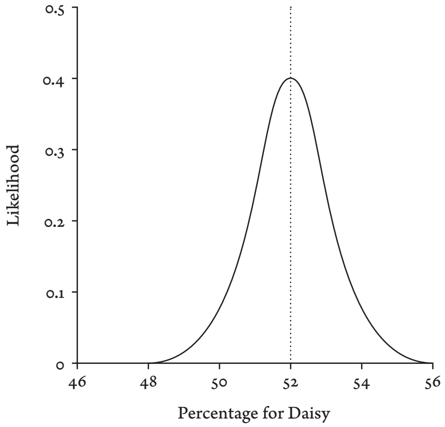

FIGURE 1 Normal distribution

Now, we are doing empirical predictions here, not mathematics, so Partial Unimodality is not a necessary truth. And Barnett himself acknowledges the possibility of cooked-up cases where Partial Unimodality fails (his example is one where one relies on two polls, suspecting that one of them is fraudulent).[11] We want to now suggest that the problem is more serious than that and that cases where Partial Unimodality fails need not be all that cooked up. Consider, then, the following examples. In all of them, Daisy is still projected to get 52 percent of the vote.

> *Systematic Mistake*: There is a part of the voting population that is tricky to capture in polls. In all likelihood, either the pollsters overcame this problem, in which case there will be no systematic error here, or they did not, in which case a rather chunky mistake is present. Perhaps, if the pollsters did not overcome the problem, Daisy is likely to get around 48 percent of the vote. In such a case, a 48 percent outcome for Daisy may be more likely than a 50 percent one.
>
> *Last-Minute Event*: The polls do a very good job at reflecting the voting plans of most at the time of conducting the poll, but some last-minute event may bring about a change in the vote of some 4 percent of the population from Daisy to her rival (Donald). If, for instance, Daisy is perceived as the more dovish candidate, perhaps a last-minute terrorist attack may have such an effect. If there is no terrorist attack, the outcome is likely to be very close to 52 percent for Daisy. If there is a terrorist

[11] Barnett, "Why You Should Vote to Change the Outcome," n27.

> attack, though, it is likely to be around 48 percent for Daisy. But around 50 percent is an unlikely result on either scenario.
>
> *Guru*: About 4 percent of the voting population will vote according to what the Guru will tell them. Right now, the Guru tells them to vote for Daisy. But he may change his mind. If he does, this will bring about a "chunky" change in voting—rendering 48 percent a more likely outcome than 50 percent.[12]

It may be argued that even if we are right about these cases, a restricted version of the Chances condition would still hold—not for very competitive elections, but for those in which the expected error in the polls (in our examples, 4 percent) is approximately equal to the leading candidate's advantage in the polls.[13] We are not sure how many real-world cases will survive this restriction.[14] In addition, there are plausible scenarios where the expected error of the poll cannot match the advantage of the leading candidate:

> *Close-Call Incentive*: Donald represents a suppressed minority that traditionally has low election turnout. Given that the latest polls predict that Donald is at least 48 percent likely to win, this very fact—the polls' projections—gives a weighty incentive for members of Donald's minority to vote. Such influence may predictably lead to Donald getting 52 percent of the vote and may make a 52 percent outcome likelier than a 50 percent one.

In this case, if the leading candidate has a large advantage, the Close-Call Incentive will not be activated, so a large error can only occur when none of the candidates has a significant edge. Therefore, there cannot be a situation where the anticipated error is close to the advantage of the leading candidate.

None of these cases, it seems to us, is too far-fetched. Indeed, the first three cases are loosely based on plausible descriptions of real elections where we come from. And as we show in a brief appendix, all cases find support in the empirical literature. When our predictions for the most likely election outcome

[12] We thank Dor Mitz for this example.

[13] Barnett suggested this response in correspondence.

[14] Remember that the Chances condition also requires that the likelihood that the leading candidate comes within ten percentage points of her opponent is at least one-half. In the case described in the text, where the elections are less competitive, this requirement is more restricting.

have a risk of systematic failure, Partial Unimodality may fail. In these cases, the likelihood distribution of outcomes looks more like it does in figure 2.[15]

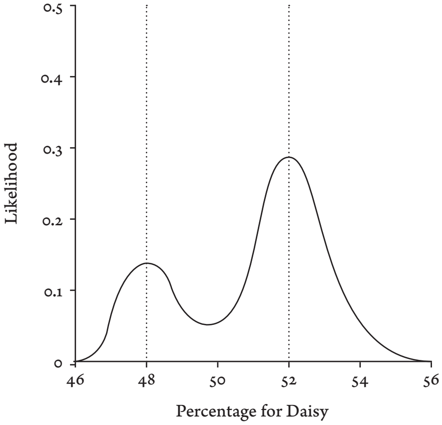

FIGURE 2 A violation of Partial Unimodality

What can be said about the conditions in which Partial Unimodality stands (or falls)? For one, if we can assume that each of the votes is probabilistically independent of any other (i.e., a premise of Independence), then we are left with a normal distribution of likelihoods around the projected results, as in figure 1, and then, of course, Partial Unimodality holds. But Independence is a very strong premise (and a highly implausible one empirically), and it is one that we have reason to believe that Barnett rejects (because he rejects the binomial model—see below). So it is important to note that he does not need Independence. He can settle for weaker premises that will nonetheless guarantee that there are no local maxima on the distribution, no "hills" of the kind that appear in figure 2 around the 48 percent line.

This, while not as strong a premise as Independence, still amounts to a highly nontrivial empirical hypothesis. As the (not-too-cooked-up) examples above show, there are quite realistic scenarios in which the no-hills hypothesis is false and, furthermore, knowably false. But, of course, in order to make more progress on Partial Unimodality, what is needed is not more *a priori* reflection, but empirical analysis.[16] In the appendix, we provide evidence that there

[15] There may be other types of realistic cases where Partial Unimodality fails, cases with a very different graph as well.

[16] *A priori* speculation can get us some of the way there, of course. We hope that the (hypothetical, if actual-world-inspired) examples above are not entirely useless. But it is not remotely enough. Perhaps, for instance (as Barnett suggested in correspondence), the effects present in Systematic Mistake, Last-Minute Event, and Guru are likely to be rather

may be instances where partial unimodality fails. However, we do not provide evidence regarding the frequency of such occurrences. We are not aware of any empirical study that is directly focused on the likelihood of Partial Unimodality. The one influential study we did find seems to indicate failures of Partial Unimodality.[17] That study is limited—partly because it focuses on very close elections, and there are not sufficiently many of those to support strong conclusions—but until stronger empirical analyses are presented, the bottom line remains the same: cases in which Partial Unimodality fails are not too far-fetched, and the speculation that they are quite common is plausible enough to pose a problem for Barnett's argument.

To the extent that Barnett's is the best case for a vindication of voting to change the outcomes, much more work needs to be done before this vindication is complete.

3. Concluding Observations

We want to conclude with several brief observations.

First, failures of Partial Unimodality may be interestingly distributed. For instance, they may not be distributed symmetrically—perhaps, for instance, Guru-like cases are more likely among Daisy's voters, or perhaps systematic mistake is more likely among Donald's. This may result in different verdicts for different voters regarding whether or not they have a reason to vote to change the outcomes. We take this to be a plausible result.

Second, there is an interesting relation between the problem with Partial Unimodality (and therefore also with Barnett's argument for the Chances condition and, with it, his argument for the conclusion that we very often have a reason to vote to change the outcomes) and Barnett's own criticism of Brennan's use of the binomial model in generating his (Brennan's) overwhelmingly small expected value for voting ("Under a binomial model, an *N*-voter election is modeled as *N* tosses of a biased coin, where the coin's bias is fixed by the specifics of the case.")[18] Barnett does not explain *what* is wrong with the binomial model, but he does give reasons—conclusive reasons, we think—to believe *that* something is wrong with it as an attempt to model real-world voting.[19] We do not need all the details here, but it is safe to say that the problem with the

rare and small, and perhaps to (pretty much) cancel each other out when present. These too are legitimate speculations (though we do not find them plausible). But we need empirical analysis to decide.

[17] Mulligan and Hunter, "The Empirical Frequency of a Pivotal Vote."

[18] Barnett, "Why You Should Vote to Change the Outcome," 431.

[19] Barnett, "Why You Should Vote to Change the Outcome," 440.

binomial model is that the distribution of likelihoods it predicts is clustered—as a normal distribution or something very close to it—around the projected outcome. In fact, the binomial model *does* presuppose Independence, and it is a plausible hypothesis that the binomial model's failure is due precisely to the empirical implausibility of Independence. But even if Independence is not the whole story of the binomial model's failure, still it is clear that something in the vicinity is—the fact that (for instance) Brennan assumes that the distribution of likelihood of outcomes clusters nicely around the projected outcome is what spells the model's doom.[20]

However, Barnett does not settle for merely showing that Brennan's model is unrealistic. He also puts forward a model seemingly showing a reason to vote to change the outcome. So, while Barnett is correct to assert that Brennan's assumption of voter independence is unrealistic, our criticism of Barnett's use of Partial Unimodality—somewhat ironically—shows that Barnett, too, falls prey to a rather similar (if less acute and conclusive) flaw.[21]

Last, we want to tentatively suggest a methodological point, for Barnett may respond by insisting that even if Partial Unimodality often fails as a matter of objective reality, still voters are rarely if ever in a position to know this.[22] As the examples above and the empirical data in the appendix indicate, this may not be so, but let us suppose that it is. Seeing that the mission Barnett has embarked on is precisely to show that it often makes sense—from a voter's perspective—to vote in order to change the outcomes, unknown "hills" in the distribution of likelihood of outcomes do not seem to matter. So, on the assumption that known "hills" are very rare, our objection to Barnett's argument seems to fail. Now, this line of thought is surely right when emphasizing that the nature of the mission here is not one that allows talk of the objective "ought" or some such. (Presumably, whether I objectively ought to vote in order to change the outcomes simply depends on whether or not, as a matter of objective fact, I end up casting the deciding vote.) So it is very tempting to add the uncertainty about potential failures of Partial Unimodality to the general uncertainty mix. But—and here is our tentative suggestion—we are not sure this is so. The

[20] In correspondence, Barnett suggested that the issues with the binomial model might be more complicated.

[21] As already briefly noted, Barnett explicitly says that he does not want to rely on any elaborate model (let alone the binomial one), and instead hopes to rely solely on specific highly plausible premises ("Why You Should Vote to Change the Outcome," 441). So it bears reemphasizing that his premises—or anyway, Partial Unimodality—smuggle back in one of the main causes for concern regarding the binomial model.

[22] In correspondence, Barnett suggested this response. The wording in the text here is sloppy for a reason we return to shortly.

uncertainty about whether or not one will cast the deciding vote is the uncertainty that defines the problem and, indeed, the mission Barnett has embarked on. Uncertainty about Partial Unimodality—that is, about possible "hills" in the outcome-likelihood distribution—seems to be of a different kind, perhaps because it is second-order (being already *about* likelihoods, presumably understood subjectively). Furthermore, such second-order uncertainty may have unique characteristics. Perhaps, for instance, while there is some reason to think that the possibility of "hills" in different places and of different "heights" along the distribution can be safely ignored when drawing conclusions about elections in general, still in many real-world cases the specific voter will have much richer, more specific information about the specific election they are facing (those in their state, say, or in their county), such that in that specific case possible hills cannot be safely neglected. So if the only way of saving Barnett's argument is by adding this second-order uncertainty into the usual uncertainty mix, the stakes will have been raised. (And, to repeat, we think that there are sometimes likely to be knowable hills, and furthermore, that the naked probability that there are such hills [that is, that the relevant instance of Partial Unimodality fails] in close elections may be quite high, so that the rational voter should not take Partial Unimodality for granted).[23]

Hebrew University of Jerusalem
aliron@ucsd.edu
david.enoch@mail.huji.ac.il

APPENDIX: EMPIRICAL SUPPORT FOR THE FAILURE OF PARTIAL UNIMODALITY

The purpose of this appendix is limited. We do not claim to offer a comprehensive survey of the literature here, nor do we attempt an assessment of how often it is that Partial Unimodality fails. Instead, our purpose here is to show that such failures are sometimes in place, and indeed, knowably so. This appendix shows, then, that the cases in the text are not too far-fetched and that what is needed for assessing whether, in a particular case or set of cases, there is a reason to vote to change the outcomes is further empirical research (and not just more *a priori* modeling).

[23] We thank Zach Barnett and Jason Brennan for very helpful and gracious comments in correspondence. We also thank two referees for this journal for valuable comments on this paper. For comments on earlier versions, we thank Ittay Nissan-Rozen.

Systematic Mistake cases are argued to be common in pre-election polling.[24] For example, one explanation for the inaccuracy of certain polls in failing to predict Trump's 2016 victory is the lack of adequate representation of non-college-educated white voters.[25] This error is paradigmatically a case of Systematic Mistake. And while such mistakes are often easy enough to recognize in hindsight, it is often very difficult for voters to determine—before an election—whether there is such a mistake, and in particular, whether the Systematic Mistake—if there is one—is similar in size to the advantage of the leading candidate.

Another explanation suggested by the Kennedy et al. analysis of the 2016 poll inaccuracies is a "late swing" of votes toward Trump. "Late swings" are used to explain the failure of polls in other cases as well.[26] The common explanation for this phenomenon is that late deciders are less politically anchored and, therefore, more susceptible to being influenced by campaign events.[27] People who are less anchored politically are also more susceptible to the impact of celebrity endorsements on their opinions.[28] Moreover, in more traditional societies, the support of traditional leaders can sway voters' decisions, particularly when goods are delivered in partnership with these leaders.[29] These studies suggest that both Last-Minute Event and Guru have empirical support in real-life cases.

Finally, it has also been observed that close elections can impact voters' incentives. According to Vogl, certain racial groups tend to be more enthusiastic about voting in closely contested elections. There are also other reasons that the closeness of the prediction may influence the outcomes in a biased way.[30] So, our Close-Call Inventive case is not too far-fetched either.

Given this evidence, it is no wonder that sophisticated statisticians incorporate measures to mitigate such systematic failures of polls in their models. In a teardown of his 2014 Senate Forecast model, Nate Silver stressed that his model must address this issue:

[24] Walsh, Dolfin, and DiNardo, "Lies, Damn Lies, and Pre-Election Polling."

[25] Kennedy et al., "An Evaluation of the 2016 Election Polls in the United States"; Silver, "Pollsters Probably Didn't Talk to Enough White Voters without College Degrees."

[26] Durand and Blais, "Quebec 2018."

[27] Fournier et al., "Time-of-Voting Decision and Susceptibility to Campaign Effects."

[28] Veer, Becirovic, and Martin, "If Kate Voted Conservative, Would You?"

[29] Baldwin, "Why Vote with the Chief?"; Brierley and Ofosu, "Do Chiefs' Endorsements Affect Voter Behaviour?"

[30] Vogl, "Race and the Politics of Close Elections"; Grimmer et al., "Are Close Elections Random?"

> In a number of recent elections, one party has either gained considerable ground in the closing stages of the race (as Democrats did in 2006) or the polls have had a strong overall bias toward one party or another on Election Day itself (as in 1994, 1998, and 2012).[31]

In order to mitigate this problem, Silver conducted a series of simulations in which systematic biases were randomly assigned. After implementing this solution, the final forecast ended up violating partial unimodality. Based on his model, the Republicans were most likely to finish with fifty-two seats, but they were more likely to hold forty-nine or fifty seats than fifty-one (fig. 3).[32]

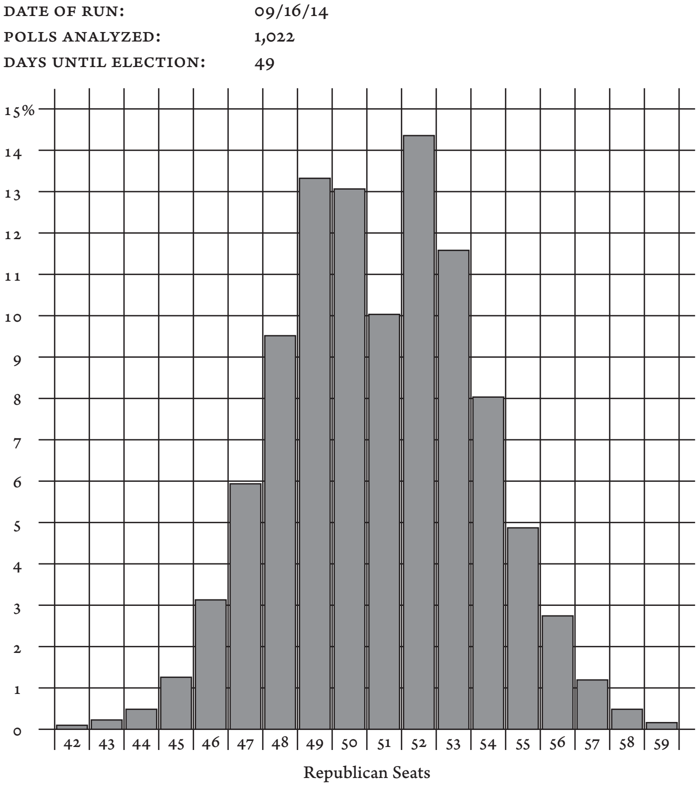

FIGURE 3 The predictions of Silver's senate forecast model
Source: Silver, "How the FiveThirtyEight Senate Forecast Model Works"

[31] Silver, "How the FiveThirtyEight Senate Forecast Model Works."

[32] We thank an anonymous referee for suggesting that we address Silver's models here. We were happy to find out that at the end of the day, Silver's models—and even graphs—support our main point in this paper.

REFERENCES

Baldwin, Kate. "Why Vote with the Chief? Political Connections and Public Goods Provision in Zambia." *American Journal of Political Science* 57, no. 4 (October 2013): 794–809.

Barnett, Zach. "Why You Should Vote to Change the Outcome." *Philosophy and Public Affairs* 48, no. 4 (Fall 2020): 422–46.

Brennan, Jason. "The Ethics and Rationality of Voting." *Stanford Encyclopedia of Philosophy*, (Winter 2020). https://plato.stanford.edu/archives/win2020/entries/voting/.

Brennan, Jason, and Christopher Freiman. "Why Swing-State Voting Is Not Effective Altruism: The Bad News about the Good News about Voting." *Journal of Political Philosophy* 31, no. 1 (March 2023): 60–79.

Brierley, Sarah, and George Ofosu. "Do Chiefs' Endorsements Affect Voter Behaviour?" International Growth Centre, 2021.

Durand, Claire, and André Blais. "Quebec 2018: A Failure of the Polls?" *Canadian Journal of Political Science* 53, no. 1 (March 2020): 133–50.

Fournier, Patrick, André Blais, Richard Nadeau, Elisabeth Gidengil, and Neil Nevitte. "Time-of-Voting Decision and Susceptibility to Campaign Effects." *Electoral Studies* 23, no. 4 (December 2004): 661–81.

Grimmer, Justin, Eitan Hersh, Brian Feinstein, and Daniel Carpenter. "Are Close Elections Random?" 2011. http://web.stanford.edu/~jgrimmer/CEF.pdf.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D. Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers, et al. "An Evaluation of the 2016 Election Polls in the United States." *Public Opinion Quarterly* 82, no. 1 (Spring 2018): 1–33.

Mulligan, Casey B., and Charles G. Hunter. "The Empirical Frequency of a Pivotal Vote." *Public Choice* 116, nos. 1–2 (July 2003): 31–54.

Silver, Nate. "How the FiveThirtyEight Senate Forecast Model Works." FiveThirtyEight.com, September 17, 2014. https://fivethirtyeight.com/features/how-the-fivethirtyeight-senate-forecast-model-works/.

———. "Pollsters Probably Didn't Talk to Enough White Voters without College Degrees." FiveThirtyEight.com, December 1, 2016. https://fivethirtyeight.com/features/pollsters-probably-didnt-talk-to-enough-white-voters-without-college-degrees/.

Veer, Ekant, Ilda Becirovic, and Brett A. S. Martin. "If Kate Voted Conservative, Would You? The Role of Celebrity Endorsements in Political Party Advertising." *European Journal of Marketing* 44, nos. 3–4 (April 6, 2010): 436–50.

Vogl, Tom S. "Race and the Politics of Close Elections." *Journal of Public*

Economics 109 (January 2014): 101–13.

Walsh, Elias, Sarah Dolfin, and John DiNardo. "Lies, Damn Lies, and Pre-Election Polling." *American Economic Review* 99, no. 2 (May 2009): 316–22.