THREE PAPERS IN ETHICS

"THE DIRECT HARM OF TESTIMONIAL INJUSTICE," "HYPOCRISY AS SELFISH SELF-EXCEPTIONALISM," AND "THE PROBLEM OF SELF-SACRIFICE."

A Dissertation

Presented to the Faculty of the Graduate School

Of Cornell University

in Partial fulfillment of the Requirements for the Degree of

Doctor of Philosophy of Philosophy

By

Avi Appel

August 2021

THREE PAPERS IN ETHICS

Avi Appel, Ph.D.

Cornell University 2021

The Direct Harm of Testimonial Injustice

What kind of wrong is it to prejudicially doubt a person's credibility or honesty? I will argue that prejudicial incredulity is immediately and unjustly harmful; in particular, I argue that the attitude of incredulity is a type of disesteem, that disesteem is directly harmful, and that prejudicially caused harms are procedurally unjust.

Hypocrisy as Selfish Self-Exceptionalism

What is hypocrisy? Why is it bad? How bad is it? I propose that hypocrisy is selfish self-exceptionalism, which is when a person holds others to a more demanding standard than that to which she holds herself due to a selfish bias. This theory explains what hypocrisy is. I also argue that the intrinsic selfishness of hypocrisy, in the context of committing to normative standards, explains why hypocrisy is bad.

The Problem of Self-Sacrifice

Decisions that are, intuitively, self-sacrificial seem not to be accommodated by unrestricted wellbeing-subjectivism. In response, proponents of wellbeing-subjectivism have proposed various restrictions or reinterpretations of the concept of self-sacrifice. I dispute each of these proposals and propose my own analysis of self-sacrifice, one that is compatible with unrestricted wellbeing-subjectivism after all.

BIOGRAPHICAL SKETCH

Avi Appel earned a BA with honors in philosophy at the University of Wisconsin - Madison, in 2009. He completed an MS in Accounting at the Boston College Carroll School of Management, in 2010. He completed a non-terminal MA in philosophy at Cornell University in 2017.

This dissertation is dedicated to all the prejudicially incredulous, the hypocritical, and the self-sacrificial, without at least some of whom (*de dicto*), this dissertation would not have been coherent.

ACKNOWLEDGEMENTS

This dissertation was greatly helped by my committee chair, Julia Markovits, who was always supportive, insightful, and available, and was even disposed to take my side in disagreements with other professors. I also received much supportive feedback from Kate Manne, Richard Miller, David Sobel, and Nicole Hassoun.

TABLE OF CONTENTS

The Direct Harm of Testimonial Injustice

Section 1 - Introduction

What kind of wrong is it to prejudicially doubt a person's credibility or honesty? Consider the following example:

> *Library Patron*: A library patron peruses the baseball section, passing over a book written by a woman because, he thinks, "what do women know about baseball?"[1] He instead selects a book written by a man. He reads the book only for his own edification; neither author receives any more or less money or fame from his action.[2]

The patron has done something unethical, intuitively, but it is puzzling how that intuition can be supported given that there are no expected (or even actual) negative downstream consequences to anyone. In resolving this puzzle, I will argue that prejudicial incredulity is immediately and unjustly harmful; in particular I argue that the attitude of incredulity is a type of disesteem, that disesteem is directly harmful, and that prejudicially caused harms are procedurally unjust.

There has been an explosion of essays on testimonial injustice in the past decade, following Miranda Fricker's 2007 book *Epistemic Injustice*[3] and the vast majority of these essays endorse Fricker's theory of the direct harm of testimonial injustice (according to which testimonial injustice directly harms a speaker by thwarting their attempt to contribute knowledge), though few explore the theory in detail. Alternative theories of the direct harm have also been offered by Shannon Sullivan (Sullivan 2017) and Gaile Pohlhaus Jr. (Pohlhaus Jr. 2014).

---

[1] This is a real stereotype; The New Yorker reports: "In 1979, [Jane] Leavy joined the sports desk of the Washington Post, becoming part of a small but growing cadre of women in the business. When women were finally allowed into the Yankees' locker room, after the 1978 court ruling, players insinuated that they were there 'to look.' 'I used legal pads and Sharpies to take notes, as a strategic shield, so I could put the large pad in front of my eyes,' Leavy said." (Helfand 2018)

[2] Another way to vividly illustrate the distinction between downstream harms and direct harms is to imagine that the victim dies immediately after the testimonial injustice has occurred. Had the victim *already* been harmed?

[3] Although Fricker coined the term "testimonial injustice" and is responsible for the upsurge in analytic philosophy on the topic, I do not mean to suggest that Fricker was the first to explore the concept of prejudicial incredulity itself. Others, especially black women, have been writing about the concept long before Fricker (McKinnon 2016, 438–39).

Why does it matter whether testimonial injustice is directly harmful? Although I grant that the downstream harms of testimonial injustice will usually be more important, there are a few reasons for spilling ink in search of a unified theory of the direct harm.

First, there is often a tight connection between injustice and harm, since injustice typically takes the form of unjustly distributed harms (or benefits).[4] A satisfactory theory of the direct harm of testimonial injustice is therefore important for supporting the intuition that testimonial injustice counts as an injustice even in those instances in which no downstream harms result. One hitch in this distributive-injustice-based motivation for theorizing the direct harm of testimonial injustice is that Fricker argues that testimonial injustice is not a distributive injustice. I disagree with Fricker on this count and explain why in section 3.1.

Second, beyond securing the intuition that testimonial injustice is unjust, getting clear about what kind of injustice and of direct harm occurs has implications for whether credibility *excesses* are unjust. (A credibility excess is when you are *overly* credulous of a speaker.) Whether credibility excesses are directly harmful, and whether they are directly unjust, is a contested question: Fricker thinks not on both counts (Fricker 2007); Emmalon Davis thinks so on both counts (Davis 2016); and Jose Medina thinks they are characteristically unjust, though only indirectly harmful (Medina 2011). In contrast, my view is that, since credibility excesses in *certain* domains (e.g. knowing how to sell illegal drugs) *can* be a form of disesteem, credibility excess is sometimes (though not characteristically) directly harmful and unjust. Unfortunately, I will not have space in this essay to further explore the harm / injustice of credibility excesses.

Third, testimonial injustice is sometimes considered to be a distinctive (and novel) normative kind *because* the direct harm involved is distinctively epistemic, whereas the default kinds of harms are moral or prudential. (A person is harmed epistemically iff harmed qua knower.) Which kinds of actions cause which kinds of harms is of some theoretical interest. Each of the extant theories of the direct harm of testimonial injustice support the claim that this direct harm is distinctively epistemic. If we reject those theories, as I argue we should, then what is the status of the "distinctively epistemic" claim? The upshot of my esteem-based theory is that the direct harm of testimonial injustice is not distinctively epistemic, since disesteem does not harm subjects *qua* knowers.

After explaining testimonial injustice (section 2) in more detail, I explain what kind of injustice it is (section 3.1), why incredulity is a type of disesteem (section 3.2), and I provide some reasons that disesteem is directly harmful (section 3.3). I then review and criticize Fricker's theory of the direct harm

[4] Indeed, Anderson has suggested an even tighter connection between the two, arguing that "there can be no injustice without an injury to someone's interests" (Anderson 2010, 5). (I set aside the alleged distinction between "injury to someone's interests" and "harms." Cf. (Gardner Forthcoming).)

of testimonial injustice (section 4). Fricker's theory claims that testimonial injustice thwarts one's attempt to contribute knowledge, thereby undermining one aspect of a person's capacity to reason, thereby dehumanizing the speaker. Alternatives to this theory have been proposed: Sullivan's claims that testimonial injustice directly harms a speaker by preventing them from co-constructing knowledge (Sullivan 2017) and Pohlhaus Jr.'s claims that testimonial injustice directly harms a speaker by otherizing / epistemically derivatizing her (Pohlhaus Jr. 2014). I also find these alternative theories implausible, but, to avoid an overly long paper – and because neither of these theories has received uptake in the literature – I do not discuss them further.

Section 2 – Describing Testimonial Injustice

What is "testimonial injustice?" The very brief definition is: a credibility deficit caused by prejudice. Before fleshing this definition out in more detail, it will be helpful to start with a couple examples. First, we have a real-life example from an interview Fricker conducted and reported on:

> *Egyptian Business Woman:* An Egyptian business woman with demonstrably useful knowledge to contribute, found that her ideas received no uptake from most of her colleagues unless she issued them indirectly through a sympathetic male colleague (e.g. by passing him notes during a meeting). (Fricker 2007, 47).

Second, an example observed by Linda Alcoff:

> *African American Academics:* "When writers from oppressed races and nationalities have insisted that all writing is political the claim has been dismissed as foolish, or grounded in ressentiment, or it is simply ignored; when prestigious European philosophers say that all writing is political it is taken up as a new and original "truth" (Judith Wilson calls this "the intellectual equivalent of the 'cover record.' (Wilson 1991, 122))"[5] (Alcoff 1991, 13)

[5] Though Alcoff emphasizes prejudices against the style in which the thesis was argued for, I imagine that she would also grant the relevance of racial credibility prejudices.

These two examples are particularly clean, since we can be sure that the plausibility of the communicated *content* is not the reason the speaker is disbelieved.[6] In each of these examples the content *was* believed once it was issued by someone viewed as being more credible.[7]

Fricker also discusses the phenomenon of speakers not even being questioned by prejudiced hearers - a phenomenon Fricker refers to as "pre-emptive testimonial injustice" (Fricker 2007, 130). Fricker's conceptual structure can be used to shed light on the concept of mansplaining and doing so serves nicely to further illustrate the concept of testimonial injustice. Mansplaining relies on pre-emptive credibility deficits. The mansplainer not only declines to question his interlocutor regarding some topic, due to a prejudicial credibility-deficit, but even goes so far as to explain the topic to her. Mansplaining, then, is kind of a pre-emptive testimonial injustice "plus one."

The longer, more nuanced definition of testimonial injustice is:

> *Testimonial Injustice*$_{def}$: when an agent, the "hearer," assigns a credibility deficit to another, the "speaker," due to one of the agent's systematic, negative, identity-prejudicial stereotypes.

I put "hearer" and "speaker" in quotes because testimonial injustice can occur even in the absence of anyone speaking or hearing anything at all. If ever a colleague of the Egyptian businesswoman, for example, were to ask some other (male) colleague for business-related advice instead of the woman (*because* of a prejudicial credibility deficit) the businesswoman would suffer a testimonial injustice even though she did not speak. But for stylistic convenience, I will often call the victim and perpetrator "speaker" and "hearer" respectively.

---

[6] See Karen Jones on the "double disadvantage" that "testifiers who belong to 'suspect' social groups and who are bearers of strange tales" suffer (Jones 2001). Arguably, it is the latter type of disadvantage that faces Tom Robinson, in Harper Lee's *To Kill a Mockingbird* (Lee 2002), when he tries to testify to an all-white jury that Maella's injuries were caused by her (white) father rather than a black man (Pohlhaus Jr. 2012, 726). Cf. (Fricker 2007).

[7] Of course, testimonial injustice is wholly compatible with incredulity based also on content; the point I am making here is merely that such examples will be less illustrative of the core concept.

It is important to note that the stereotype must be negatively valenced. In the following scenario, Beth is prejudicially incredulous of white people, but does not do them an injustice *because* the credibility deficit involved is not negatively valenced:

> *Narcotics Officer:* Beth, a police officer, wants to find out where illegal drugs are sold. She sees several people hanging out in the part of town where junkies are known to hang. Because of a prejudicial stereotype, Beth believes that white people are unlikely to know where to buy drugs. So she passes over the white people she sees, directing her inquiries only at non-whites.

Intuitively, Beth commits an injustice against the non-white people by *overestimating* their credibility in the negatively-valenced domain of narcotics knowledge. Fricker's theory of testimonial injustice fails to capture this intuition, since credibility excesses cannot be directly unjust on her theory. I suggest the following unfriendly amendment to Fricker's theory:

> *Testimonial Injustice*$_{def}$: when an agent, the "hearer," assigns a credibility deficit (*or excess*) to another, the "speaker," due to one of the agent's systematic, negative, identity-prejudicial stereotypes.

As I will argue in the next section, the negative valence of the stereotype does enough normative lifting to obviate the need to restrict the credibility mistakes to deficits only.

One of the more confusing and controversial parts of the definition is what it takes for an identity-stereotype to be prejudicial. What it takes for an attitude to arise from prejudice – and to what extent prejudices are blameworthy - will be relevant when I defend my theory of the injustice of testimonial injustice, since some have the intuition that blameworthy actions are a necessary condition for injustice. I will address the prejudice debate now.

In *Epistemic Injustice*, Fricker defined prejudice as "judgments [that] may have a positive or negative valence, and [that] display some epistemically culpable resistance to counter-evidence owing to some affective investment on the part of the subject" (Fricker 2007, 35). Note that, on this definition, prejudice is *defined* as culpable.

But several philosophers have taken issue with Fricker's original characterization of prejudice (e.g. (Alcoff 2010), (Maitra 2010), (Riggs 2012), (Begby 2013), (Battaly 2017)) and in more recent work, Fricker has updated her understanding of prejudice to include any attitude that "is the product of (some

significant degree of) motivated maladjustment to the evidence" (Fricker 2016, 38).[8] An affectively driven disposition to resist counter-evidence, on this updated view, is just *one way of several* to have a motivated maladjustment to evidence.

Fricker has also updated her understanding of culpability's correlation with prejudice: prejudices no longer *require* culpability; although prejudices do require some sort of pseudo-culpability at a minimum. What does this pseudo-culpability amount to? To illustrate the kind of pseudo-culpability involved (which Fricker refers to as "non-culpable fault-worthiness"), Fricker describes a case in which she thinks a non-culpable prejudice can occur:

> "If, for example, someone chairing an academic appointments process were systematically, albeit unwittingly, to assess the male candidates' writing samples more highly than those of the female candidates owing to the operation of implicit prejudice in her patterns of judgement, then other things equal we would regard her as epistemically at fault, and so blameworthy.
>
> We might not blame her very much, of course, if she were doing her well-intentioned best under difficult circumstances—pressure of time, lack of institutional support for alternative methods. These are mitigating circumstances, or excuses, and they function to reduce the degree of appropriate blame, even to zero in some cases, if we accept the possibility of fully exculpatory excuses, which we surely may. But they do not change the kind of epistemic fault that has expressed itself, which is blameworthy other things equal." (Fricker 2016, 39–40)

Fricker goes on to claim that we could imagine that the writing sample assessor was not herself making any motivated maladjustment to the evidence but was instead a "conduit" for prejudices inherited from her environment. Given the "*environmental epistemic bad luck*" (Fricker 2016,

[8] David Sobel has suggested a counter-example to this motivated-maladjustment definition of "prejudice" (in conversation): suppose a doctor diagnoses me with cancer, but I do not believe him because of a motivated maladjustment to the evidence. The motivation will be some sort of bias towards oneself (i.e. "I am a healthy person. I can't have cancer!). I do not think this problem in the definition of prejudice infects the larger definition of testimonial injustice, however, since testimonial injustice is concerned only with *systematic identity*-prejudices.

45, no emphasis added), "she would appropriately experience her responsibility for what she has done in the mode of *epistemic agent-regret*" (Fricker 2016, 47, no emphasis added).[9] Though I am not endorsing Fricker's analysis of her academic search committee example[10], I do agree that there will be some cases of prejudice that are non-culpably learned from others (i.e. some cases of pernicious associations that are not influenced by the believer's *own* blameworthy motivations).

To summarize, testimonial injustice is negatively valenced, (social-identity-) prejudicial incredulity; and Fricker quite plausibly proposes that "prejudice" requires, at a minimum, some sort of epistemic agent-regret. So far so good; in the next section I turn to the normative evaluation of testimonial injustice, where the disagreement begins.

Section 3 – Evaluating Testimonial Injustice

What kind of wrong is testimonial injustice? My answer comes in two parts. First, I argue that the kind of injustice involved is an unjust distribution of harms. Second, I argue that the harm unjustly distributed by testimonial injustice is disesteem. These two parts together suffice to explain why testimonial injustice is inherently wrong.

Section 3.1 – The Injustice of Testimonial Injustice

[9] See also theses along similar lines from Jeremy Wanderer ("…it is not incoherent to talk of the harm caused in treating someone as an object through nonattribution of testimonial status without accusing the agent of actively perpetrating the injustice") (Wanderer 2012, 165), (Anderson 2012), and (Battaly 2017).

[10] As Nicole Hassoun pointed out to me (in conversation), surely the minimal epistemic / moral standards for academic search committee members include examining their own prejudices against women applicants.

What kind of injustice is testimonial injustice? Let us start with Aristotle's observation that

> "One kind [of justice] is that which is manifested in distributions of honor or money or...other things…[A]ll men agree that what is just in distribution must be according to merit in some sense...The just, then, is a species of the proportionate...the unjust is what violates the proportion." (Aristotle 1941, Book V: Chapters 2-3, lines 1131a1-1131b17)

The following proposal, a *sufficient* condition of injustice, borrows Aristotle's insight that the unjust violates the proportion that ought to exist between what a person merits and what they receive:

> **Prejudicial Procedural Distributive Injustice**: An award of benefits (or harms) is unjust to the extent that the amount the recipient(s) receives is influenced by identity-prejudice.[11]

This principle posits a "procedural" injustice because of its focus on the upstream causes of the distribution. Procedural injustice is typically contrasted with "substantive" injustice, that is, an injustice in the outcomes. Testimonial injustice is always procedurally unjust, but not always substantively unjust. Take, for example, a situation in which a woman lies about how qualified she is. Her hearers give her less credulity *qua* woman, but also more credulity *qua* perceiving her as an expert. If these two credibility factors balance out, then she may well receive precisely the amount of *net* credibility that is warranted. The woman has suffered a procedural injustice, since others assigned her less credibility than they would have in the absence of their sexist stereotypes, but not a substantive injustice, since she received precisely the amount of credulity that she merited.

I think the proposed principle is fairly intuitive on its face, but one worry we might have is whether it obeys a constraint on "judgments of injustice" – proposed by Elizabeth Anderson – which says that "there can be no injustice without an agent who is (or was) substantively responsible for it – someone obligated to avoid, correct, or bear the costs of the injustice or of its correction or amelioration" (Anderson 2010, 5). Anderson motivates this constraint with an example from the movie *Amadeus*, in

[11] By relying on a merely sufficient condition of injustice, my definition does not rule out the existence of unsystematic and non-prejudicial cases of injustice such as when an innocent, upper-class, white man is found guilty due to a fluky bureaucratic incompetence on the part of the judicial system. Non-desert-based kinds of injustices may also exist. Rather than try to unify all kinds of injustice under one principle (which may, for all I know, be possible), it suffices for my purposes to identify a single kind.

which Salieri is jealous of Mozart for the latter's unearned, superior musical talent. In Salieri's world-view, natural talent should be distributed in proportion to virtue, and thus, as Anderson notes, Salieri addresses his complaint to God. The point is that a complaint of injustice must be addressed to *someone responsible* (Anderson 2010, 9–10). But if prejudices can occur in the absence of blame (as explained in the previous section), then it seems that Prejudicial Procedural Distributive Injustices could occur without someone responsible to address our complaint to.

For a more credibility-relevant version of Anderson's *Amadeus* example, imagine a dolt who nobody deems credible, not because of prejudice but because he is demonstrably stupid. (Imagine further that the state has made every reasonable effort to educate him and ameliorate his intelligence.) One might plausibly argue that there is something *unfair* in his natural lack of intelligence. I am sympathetic to that complaint. But if Anderson is correct, then the dolt, like Salieri, could not argue that the unfairness of the genetic lottery amounted to an *injustice*.[12] And so even though the harm done to the dolt resembles the harm done to a victim of testimonial injustice, the dolt has not suffered an injustice.[13]

So the relevant question for determining whether Procedural Distributive Injustice as applied to testimonial injustice obeys Anderson's constraint, is whether a perpetrator of testimonial injustice really is "obligated to avoid, correct, or bear the costs of the injustice or of its correction or amelioration." When a testimonial injustice occurs because of a non-culpable prejudice, how can Anderson's second constraint be met? Two answers. First, an individualistic response: On Fricker's updated view of prejudice (see section 2), although some prejudices are not culpably held, they still bear the kind of pseudo-blame that we associate with agent regret. Such pseudo-blame may be sufficient to meet Anderson's constraint. Perhaps such wrongdoers are still obligated to bear the costs of correcting or ameliorating the injustice, even if it would be unreasonable to expect them to avoid perpetrating the offense.[14]

[12] The dolt might be able to make a luck egalitarian argument in favor of reparations for his lot in life, but that would not be equivalent to complaining that not being treated as competent by others counted as an injustice. It is the latter component that is relevant for my purposes.

[13] Maitra notes that other kinds of injustice besides testimonial injustice can also give rise to the harm of incredulity: e.g. in a world where blacks are not educated, a black person may be discriminated against for a job as a result of a non-prejudicial stereotype. Although the incredulity has an unjust origin, the employer's incredulity is not itself prejudicial. (Maitra 2010)

[14] Indeed, that reparations are often an appropriate consequence or expression of agent regret was one of Bernard Williams's theses regarding agent regret, in the paper in which that phrase was coined (Williams 1981, 28–29).

Second a structural response: It seems likely that some group, or institution, *is* to blame for the stereotypes the prejudiced person holds (or, in Fricker's words, "inherits"). The hearer might be a sort of "patsy" from this perspective, with some institution or group of persons, upstream from the hearer, obligated to bear the costs of preventing the injustice. Thus, Anderson's second constraint can be met albeit not in the place we first thought to look.

Third, note that this "liability" objection is not specific to *my* proposed analysis of the injustice of testimonial injustice. *Anyone* who argues that testimonial injustice is an injustice (and who maintains that prejudices can be non-culpably held) will face this objection.

And, of course, one could reject Anderson's constraint, holding that finding faultworthy agents is not a necessary condition for injustice. To recap so far: I have proposed a general principle of injustice and supported the principle by explaining how it obeys one prima facie troubling constraint on principles of injustice.

The principal of injustice I proposed concerns unjust distributions of *benefits or harms*, but how does that apply to testimonial injustice? Fricker argues that distributive injustice is not relevant to testimonial injustice:

> "…credibility is not a good that belongs with the distributive model of justice. Unlike those goods that are fruitfully dealt with along distributive lines (such as wealth or health care), there is no puzzle about the fair distribution of credibility, for credibility is a concept that wears its proper distribution on its sleeve. Epistemological nuance aside, the hearer's obligation is obvious: she must match the level of credibility she attributes to her interlocutor to the evidence that he is offering the truth. Further, those goods best suited to the distributive model are so suited principally because they are finite and at least potentially in short supply." (Fricker 2007, 19–20)

Let me address each of Fricker's points. First, Fricker claims that there must be a puzzle about the fair distribution of a good in order for that good to fall under a distributive model of justice. In one sense, there is no puzzle: credulity should be distributed in proportion to whatever credulity each person merits. But every distributable good fits *that* simple model: health care, wealth, etc. should all be distributed in proportion to the amount of health care, wealth, etc. that each person merits. That is a concealed tautology. In another sense, there is a puzzle: how much credulity *does* each person merit? But that is the same kind of puzzle that exists with every other kind of distributable good as well. Fricker claims that credibility "wears its distribution on its sleeve," but I find that claim obviously false, especially when the hearer is not familiar with the speaker and has *only* stereotypes to assess the speaker's prior probability. So there is no disanalogy between credulity and other kinds of distributable goods with respect to the puzzle over who merits what.

Second, Fricker claims that distributable goods must be finite. But Procedural Distributive Injustice concerns the causes of the amount of goods received; it is neutral regarding whether those goods were "distributed" or not. The "distributive" in the name of the principal is somewhat technical. I would hesitate to say that, for example, a preschool teacher who loves one student more than another (for racist reasons, say) was "distributing" his love unfairly (since more love for one does not entail less love for the other), but I would *not* hesitate to apply my principal to this case.

So Fricker's worries about the distributive model are unfounded. Another worry one might have is that classic cases of meritocratic injustice seem to require that the distributed goods are beneficial or harmful. For example, giving one person $10 and another $5 is straightforwardly unfair (all else equal) because receiving more money is more beneficial. In contrast, giving one person a handful of sand and another person two handfuls of sand is not unfair since receiving handfuls of sand is neither beneficial nor detrimental (in other words, we are at a loss as to which recipient's shoes we would rather fill[15]). So in order to apply the meritocratic injustice model to testimonial injustice – that is, in order to establish the *unfairness* of disproportionate incredulity – we need an explanation for why credulity is beneficial and incredulity detrimental. I will address this issue in the next section, where I argue that incredulity constitutes a direct harm.

Section 3.2 – Incredulity is a Type of Disesteem

The analysis above rests on the assumption that incredulity is harmful. I argue for that premise in two parts: First, I argue that testimonial injustice is characterized by disesteem; Second, I argue that disesteem is a direct harm.

Incredulity consists in judging that a person is incompetent, ignorant, insincere, or otherwise unreliable. Each of these traits are (typically) negatively valenced and, therefore, to assign someone a credibility deficit implies that one holds them in lower esteem (all else equal).

A study of the ordinary valence of words, conducted by Amy Beth Warriner et al, confirms that incompetence, ignorance, and insincerity have negative valence. Warriner et al surveyed 1,827 people, collecting 303,539 observations on the valence of 13,915 English words, using a 1-9 "sad-to-happy" scale for each word. Results relevant for my arguments are excerpted below (Warriner, Kuperman, and Brysbaert 2013):

[15] With our feet – not with the sand.

Word	Average Valence (on 1-9 scale)
knowledgeable	7.95
ignorant	3.24
reliable	7.30
unreliable	2.74
trustworthy	7.25
untrustworthy	2.67
truthful	7.48
dishonest	3.00
competent	6.05
incompetent	2.77

None of these results strike me as even remotely surprising. Testimonial injustice, therefore, involves believing (implicitly or explicitly) that the speaker has one or more negatively valenced characteristics. Believing that someone has negatively valenced characteristics *just is* to disesteem that person (pro tanto).[16]

Other philosophers writing on testimonial injustice have noted in passing its connection to disesteem (Congdon 2017, 248) and close cousins thereof, such as dishonor (Fricker 2007, 46) and disrespect (Origgi and Ciranna 2017, 305) (Congdon 2017, 249). That there is *some* connection between disesteem and testimonial injustice is not controversial. My theory is novel in its claim that the connection between disesteem and testimonial injustice is tight enough to fully ground an explanation of the direct harm of testimonial injustice.

One might wonder what my theory has to say about cases where a credibility deficit has *positive* valence. In such cases (all else equal) no disesteem occurs; thus my analysis says nothing. That is not a drawback, however. Recall *Narcotics Officer*, from section 2: if a credibility deficit has positive valence, it

[16] My definition of esteem involves a belief that someone has a characteristic towards which one takes a negative attitude. Alessandra Tanesini's definition is, I believe, essentially the same: "[Esteem is] a positive or negative attitude, directed at a person, group or institution for their good or bad qualities." (Tanesini 2018, 49)

does not count as testimonial injustice (by definition). So my theory need not say anything about such cases given that my project is to evaluate testimonial injustice. Apart from my analysis of testimonial injustice, however, a positively valenced judgment of a person is a way of esteeming them, which ought directly to benefit them. The only reason this is surprising is because doubting someone's credibility is *usually* a way of disesteeming them. (So the case is analogous to the normative status of "throwing someone into the briar patch:" it is usually, but only contingently, harmful.)

Section 3.3 – The Direct Harm of Disesteem

So far, I have argued that one unfairly distributes esteem if one's assignment of credibility deficits is influenced by identity-prejudice. However, such unfairness will only count as a procedural distributive injustice if disesteem is a kind of harm. (If disesteem is not harmful, who cares how it's distributed?) To support my claim that disesteem *is* harmful, I will appeal directly to intuitions regarding several cases from fiction and reality. In each of these cases, I invite the reader to decide whether the instance of esteem seems directly beneficial / the instance of disesteem seems directly harmful. After these intuition pumps, I explain how disesteem-as-harm fits (or does not) with standard theories of wellbeing. For those, like myself, who have strong first-order intuitions about the harmfulness of disesteem, it will be a strike against those wellbeing theories that are unaccommodating.

Here are the cases:

> *The Sixth Sense*: Cole, who is able to communicate with the dead, has the following conversation with his mother, Lynn:
>
> COLE
> Grandma comes to visit me sometimes.
>
> (Lynn becomes still. Her face is unreadable. When she speaks, her words are extremely controlled.)
>
> LYNN
> Cole, that's very wrong. Grandma's gone. You know that.
>
> COLE
> I Know.
>
> (Beat.)
>
> COLE
> She wanted me to tell you– –
>
> LYNN
> (soft)
> Cole, please stop.
>
> COLE

She wanted me to tell you, she saw you dance.

(Lynn's eyes lock on Cole's.)

COLE

She said when you were little, you and her had a fight right before your dance recital. You thought she didn't come to see you dance. She did.

(Lynn brings her hand to her mouth.)

COLE

She hid in the back so you wouldn't see… She said you were like an angel.

(Lynn begins to cry.)

COLE

She said, you came to her where they buried her. Asked her a question… She said the answer is "Everyday."

(Lynn covers her face with her hands. The tears roll out through her fingers.)

COLE

(whispers)
What did you ask?

(Beat. Lynn looks at her son. She barely gets the words out.)

Lynn

(crying)
Do I make her proud?

Public Humiliation Cases: Although less common today than in the distant past, public shame as a means of official punishment is still in use. Here are two real-world examples:

1) "Part of Rebecca Escobar's punishment for killing a woman in a drunken-driving accident is repeated public humiliation. Once a month Escobar makes an agonizing hourlong trek around the county courthouse in Wilkesboro, N.C., clutching a hand-written sign. 'I am a convicted drunk driver,' the sign announces in black and red. 'And as a result I took a life.' " (Deardorff 2000)

2) As punishment for falsely claiming that they had served in the military, in combat tours in Iraq and Afghanistan, (among other offenses) a county judge in Montana "mandated that every year during the suspended portions of each of their sentences, both men must stand at the Montana Veterans Memorial for eight hours on each Memorial and Veterans Day wearing a placard that reads, 'I am a liar. I am not a veteran. I stole valor. I have dishonored all veterans.' " (Rosenbaum 2019)

The Sopranos: In S1E10 of *The Sopranos*, Tony Soprano (a mafia boss) is invited to go golfing with his next-door neighbor (a doctor) and his neighbor's friends (stockbrokers) at the neighbor's country club. Afterwards, Tony discusses the incident with his psychiatrist. Tony compares the situation he was in with that of Jimmy Smash, someone he knew as a kid. Jimmy Smash had a cleft palate and a lisp. Tony and his friends would invite Jimmy to hang out with them when they were bored. They made fun of Jimmy and laughed at him. This was why Tony's neighbor and his neighbor's friends invited Tony along – they wanted to be entertained by his mafia stories; they wanted to laugh at him.

With these cases I hope to have persuaded you of the intuitiveness of the claim that being esteemed is a benefit and being disesteemed is a harm. Lynn's mother benefits Lynn by taking pride in her, Rebecca Escobar and the false veterans are harmed by beings disesteemed by passersby reading their signs, and Tony Soprano's neighbors harm him by disesteeming him for his occupation.

Let us assume, then, that disesteem is intuitively directly harmful. Does it make sense, in theory, for disesteem to be directly harmful? An event harms a person when and because it makes that person worse off. A complete theory of harm has two components: a formal component (the event caused the person to be worse off? The person is worse off compared to before? Or worse off counterfactually?) and a substantive component (worse off on what metric?) (Gardner Forthcoming). The formal component is not relevant for assessing the harmfulness of disesteem.

The substantive component of a complete theory of harm *just is*, to simplify things, a theory of wellbeing. There are three families of wellbeing theories in the literature: subjectivist, objective list, and hedonic. Subjectivist theories claim that a person is worse off to the extent things are going against their subjective likes (or desires, values, etc,). So long as a victim of testimonial injustice dislikes being

disesteemed, subjectivist theories of desire can easily accommodate the claim that testimonial injustice directly harms[17] – no downstream consequences needed!

Objective list theories have a list of items that make a person's life better off (whether the person cares about the items or not). In all the objective lists I have come across in the wellbeing literature, none include esteem. These objective lists could be augmented, ad hoc, to include esteem. The items already on these lists strike me as ad hoc in the first place, so why not?

Hedonic theories, which claim that only positive mental states make your life better and negative mental states make your life worse, cannot accommodate the claim that disesteem is directly harmful. Recall *Library Patron*, where the victim of the testimonial injustice never finds out that she has been disesteemed. Her mental states were never impacted.

In sum, disesteem is intuitively harmful and its status as a direct harm can be accommodated by two of the three families of wellbeing theories.

Section 4 - Fricker's Account(s) of the Direct Harm

In this section I present and criticize Fricker's account of the direct harm of testimonial injustice from her book, *Epistemic Injustice* (Fricker 2007). Next, I present and criticize Fricker's account of the harm of testimonial injustice in her more recent essay, "Epistemic Contribution as a Central Human Capability" (Fricker 2015).

In *Epistemic Injustice*, Fricker argues that the direct harm of testimonial injustice comes in the form of a specific type of dehumanization of the speaker. Namely, when a speaker is excluded from contributing knowledge (as an informant) to the pool of collective understanding, they are thwarted in their capacity as a giver of knowledge, thus also in their capacity as a knower, and thus also in their capacity as a reasoner (Fricker 2007, 44–45).[18] Below I present an attempt to formalize Fricker's argument:

1. Testimonial injustice thwarts one's attempt to contribute knowledge.

[17] Recall that "directly harms" means harms even in the absence of downstream consequences. It does not mean intrinsically harms. This qualification is important, because "intrinsically" sometimes means "non-contingently," and the subjectivist theory makes *all* harms contingent on the victim disliking the event in question.

[18] By informant, Fricker means that a person is a credible and trustworthy epistemic agent. This is in contrast to being a mere source of information (objects, after all, are often sources of information).

2. Contributing knowledge is an execution of one's capacity as a knower.
3. Therefore (from 1, 2) testimonial injustice thwart's an attempt to execute one's capacity as a knower.
4. Being a knower is one aspect of the capacity to reason, which is a capacity that lends humanity its distinctive value.
5. One is Function-Dehumanized if one is thwarted in one's attempt to execute one aspect of a capacity that lends humanity its distinctive value.
6. Therefore (from 3, 4, 5) testimonial injustice is a type of Function-Dehumanization.
7. Function-Dehumanization is a direct harm.
8. Therefore (from 6, 7) testimonial injustice directly harms (because it is a type of Function-Dehumanization).

Allow me to explain "Function-Dehumanization" (premises 5 and 7) in more detail. I will explain Function-Dehumanization by contrasting it with other (more plausible) types of dehumanization; this will clear the way to reject premise 7 without undermining the normative status of dehumanization in general. First, note that there is a distinction between directly preventing someone from *having* a capacity to X, on the one hand, and undermining someone's attempt to *execute* their capacity to X. Suppose, for example, the capacity in question is getting a high GPA. A teacher can thwart a student's execution of that capacity by giving them a single low grade. But it would take much more drastic measures to prevent them from having the capacity. I use the label "Function-Dehumanization" rather than, say, "Capacity-Dehumanization," to respect the capacity-executing vs. -having distinction and to keep clear what kind of undermining Fricker is talking about.[19] Presumably, Fricker avoids claiming that testimonial injustice directly prevents someone from *having the capacity* to contribute knowledge, since that would be trivially implausible.[20]

[19] The use of "function" to refer to the execution of a capability I borrow from Amartya Sen (Sen 1985, 200).

[20] Julia Markovits has suggested, in conversation, that perhaps Fricker could argue that testimonial injustice (attaching, as it does, to systematic stereotypes), have enough of an aggregate effect to prevent some subjects from even having the capacity to contribute knowledge. Although I am sympathetic to grounding the wrong of systematic discrimination in aggregate effects, it is insufficient for Fricker's purposes. None of the examples of testimonial injustice that Fricker uses are cases in which

Second, note that Function-Dehumanization is different from Type-Dehumanization, according to which one is dehumanized if one is treated the same way one treats a non-human or object. I draw attention to this distinction because Type-Dehumanization is the "default" that comes to mind when one hears the word "dehumanization." But Type-Dehumanization is not relevant to testimonial injustice since non-humans are *not* typically assigned credibility deficits. To treat someone as being untrustworthy is not to treat them as a non-human since non-humans are not typically taken to be willful deceivers.[21] And to treat someone as being an unreliable source of testimony is not to treat them as a non-human, since non-humans are not typically treated as sources of testimony (reliable or not).[22] In support of this point, consider Matthew Congdon's example, in which a Bosnian citizen is interrogated by Americans for information, doubting his repeated testimony that he knows nothing about an embassy bombing plot because of anti-muslim prejudice. I agree with Congdon's analysis of his case:

> While multiple injustices are at work here, it is clear that testimonial injustice is among them. Yet characterizing this as epistemic objectification is misleading for at least two reasons. First, the testimonial injustice at work would not be possible unless the interrogators view Mr. B as the bearer of critical information and so treat him as a competent epistemic subject. Second, as soon as Mr. B's captors admonish him for being deceptive, they thereby include him within the sphere of potential informants to whom norms of epistemic exchange apply. Mr. B is thus treated as a subject in the dual sense of (1) being the subject of knowledge and (2) being subject to epistemic norms." (Congdon 2017, 247)[23] [24]

The purpose of this example is to help illustrate what function-dehumanization is not (it is not type-dehumanization) but also to make clear that type-dehumanization would not serve Fricker's purposes either.

---

the speaker lacks the capacity, in general, to contribute knowledge. (Even the woman in *Egyptian Businesswoman* manages to get her male colleagues to receive her knowledge, albeit by letting someone else take credit for her ideas.)

[21] I say typically, because, of course, some non-human animals can willfully deceive (camouflaging octopodes come to mind).

[22] See Jeremy Wanderer on the distinction between "ignoring the person's status as a testifier…[and] rejecting the person's status as testifier" (Wanderer 2012, 164)

[23] Gaile Pohlhaus, Jr. makes a similar point (Pohlhaus Jr. 2014, 104).

[24] See (Manne 2016) on the many ways that "dehumanizing" behavior can often only plausibly be carried out by perpetrators that see their victims as human.

Another important kind of dehumanization is what we might call Moral-Status-Dehumanization, according to which one is dehumanized if one is treated as less valuable or of lower moral status than other humans. I interpret Fricker to be arguing that Function-Dehumanization is a subcategory of Moral-Status dehumanization. She writes: "The fact that the primary injustice involves insult to someone in respect of a capacity essential to human value lends even its least harmful instances a symbolic power that adds a layer of harm on its own: the epistemic wrong bears a social meaning to the effect that the subject is less than fully human" (Fricker 2007, 44).

Although I take Moral-Status-Dehumanization seriously, I do not think Function-Dehumanization (being undermined in an attempt to execute a capacity that lends humanity its value) holds much normative weight. It is not wrong in and of itself. It may not even be typically wrong. There are too many ways to execute each of our valuable capacities for any one of those executions to be of fundamental normative significance. Consider, for example, the calculation of mathematical products via slide ruler; this is an execution of the capacity to reason. If a university declines to provide an education in the proper use of slide rulers, they are not dehumanizing students who wish to learn this obsolete skill. Another example: instead of letting customers compare relative values of food, a grocery store displays not only the price and weight of each good, but also the price-per-unit-of-weight, thereby discouraging customers from executing their capacity to divide numbers. Stores that do this do not dehumanize their customers. Thus, being thwarted in the execution of one's capacity to reason is not wrong in and of itself. The ways one executes one's capacity to reason are too broad. Of course, being denied the *capacity* to reason – that is a plausible candidate for holding fundamental normative weight. But that is not the same as being thwarted in *a particular execution* of one's capacity to reason. And one's capacity is not denied any given instance in which a person is prevented from contributing to a shared pool of knowledge.[25]

Perhaps Fricker could respond that communicating knowledge is essential to one's capacity to reason and this is what distinguishes it from mathematical calculation. Granted, learning how to reason well requires (helpful) feedback; so communicating what one is thinking is essential to being a competent reasoner in at least one respect. But the examples of testimonial injustice that are held up as paradigm in the literature are about mature, intelligent adults. No single act of communicating knowledge is *essential* to reasoning and this is especially so for someone who has already learned how to reason.

[25] One might worry that these are cases of not aiding someone in reasoning, which is less wrong than actively blocking someone from reasoning (thanks to Julia Markovits for this suggestion). The slide ruler case can be altered as follows: a group of students orders a batch of slide rules in order to learn slide rule multiplication. The University interferes and cancels the order (refunding the students their money). The University has done something wrong here, but being undermined in an attempt to execute a capacity to reason seems like a red herring. After all, it would have been just as wrong if the students were attempting some other activity, such as counting blades of grass, and the University interfered with an order of counting-tickers.

Perhaps Fricker could argue that the aggregate effect of repeated testimonial injustices prevents, or at least partially undermines, one's having the capacity to be a knower or to reason. Granted, testimonial injustices, attaching as they due to systemic identity-stereotypes, do concentrate on minority persons. But recall *Egyptian Businesswoman*. The impression one gets is that the woman was quite smarter than her colleagues. Her capacity to know and to reason was not diminished even though she was subject to systematic testimonial injustice.

So much for premise 7 then. Premise 4 claims that reason lends humanity its distinctive value. First, what exactly does this mean? It surely cannot mean that only humans reason – that is too obviously false. Fricker does not tell us much by way of elaboration. This is what she says:

> "The capacity to give knowledge to others is one side of that many-sided capacity so significant in human beings: namely, the capacity for reason. We are long familiar with the idea, played out by the history of philosophy in many variations, that our rationality is what lends humanity its distinctive value." (Fricker 2007, 44)

One such "variation," at which I assume Fricker means here to gesture, is Aristotelian Perfectionism. Thomas Hurka, a recent proponent of this school of thought, writes:

> "The most important Aristotelian claim is that humans are essentially rational…Humans are rational because they can form and act on beliefs and intentions. More specifically, they are rational because they can form and act on sophisticated beliefs and intentions, ones whose contents stretch across persons and times and that are arranged in complex hierarchies. These last features distinguish human rationality from that of lower animals." (39)

So perhaps what Fricker has in mind when she uses the word "rationality," is the kind of sophisticated rationality that Hurka describes. This would make sense of Fricker's claim that "rationality is what lends humanity its distinctive value."

But that argument is problematic. However plausible it may be that humans tend to form and act on beliefs and intentions that are more sophisticated than those of non-humans (c.f. (Hunt 2017) on octopus intelligence), it is not true that testimonial injustice requires sophisticated beliefs and intentions. A human with a severe cognitive impairment might be incapable of the kind of advanced rationality that allegedly distinguishes humans from "lower" animals. Yet this human is still capable of suffering a testimonial injustice. Presumably the kind of direct harm a testimonial injustice inflicts on this person is similar (or even equivalent) to the kind inflicted on an abled adult. But that equivalence is

unavailable to Fricker if her theory of the direct harm of testimonial injustice is built on a sophisticated capacity to reason.[26]

Perhaps Fricker could modify premise 4 by abandoning the claim that the capacity to reason lends humanity *distinctive* value. Once she drops the "distinctive" label, she can invoke a less sophisticated kind of reasoning in her argument, one shared by animals (though the word "dehumanizing" would be less apt). In one sense, this would be a step in the right direction, because then the capacity in question could be had by cognitively disabled humans. And then such humans would be capable of being directly harmed by testimonial injustice (which is the intuitive fact Fricker's original account needs to capture). Unfortunately, the capacity for minimal, unsophisticated reasoning does not have fundamental value; even a rudimentary computer can be programmed to have that kind of capacity and rudimentary computers only have non-fundamental value.

So much for premise 4, then. Premise 1 claims that testimonial injustice thwarts the transmission of knowledge. But, I shall argue, testimonial injustice often does not have this effect. In other words, testimonial injustice does not always cause epistemic exclusion. Here are five counterexamples to premise 1.

First,

> Loan Officer: Suppose a loan officer meets with two separate clients, one male and one female. Each client proposes a business plan for which each needs a $100,000 line of credit. As far as the loan officer's evidence indicates, each business has an 85% chance of success and so deserves a 5% interest rate. However, this loan officer is generally overly optimistic (i.e. he is not so great at his job!). However, due to his prejudice against the credibility of businesswomen (as such), his interpretation of the evidence is still unbalanced. He concludes that the man's business has a 95% chance of success and he concludes that the woman's business has only a 90% success chance. So he charges her a higher interest rate than him on the line of credit. Note that both clients receive more credibility than is epistemically warranted.[27]

[26] This objection inspired by Peter Singer's discussion of the superior intelligence of many non-humans to those humans with severe cognitive disabilities in "Speciesism and Moral Status" (Singer 2009).

[27] Boudewijn de Bruin has also made the connection between discriminatory interest rates (albeit in the context of mortgages) and testimonial injustice (de Bruin 2014, 15–16).

Intuitively, this is a case of testimonial injustice. And, in fact, it fits all the criteria of Fricker's definition. But the woman has *not* been epistemically excluded (has not been thwarted in her capacity to contribute knowledge). The loan officer believes all the facts (about her business) that the woman communicates to him.[28] Yet her business acumen – her credibility as an entrepreneur – is still perceived as lower due to the loan officer's sexist prejudice.

Second,

> *Persuasive Speaker:* A hearer is prejudicially incredulous of a speaker. But the speaker argues so persuasively that, in the end, the hearer believes the speaker *despite* believing, for prejudicial reasons, that the speaker is less competent (than the hearer would have believed in the absence of the prejudice). The hearer continues to doubt the speaker's credibility on future occasions (since her prejudice is resistant to counter-evidence).

In this example, testimonial injustice occurred even though knowledge *was* communicated.

Third,

> *Proofreader Interview:* Suppose Dan is interviewing candidates for a job as a proofreader. He asks each candidate some questions about grammar, spelling, etc. Dan already knows the answers to all the questions he asks. He gives harder questions to women since he thinks that women are less knowledgeable about spelling and grammar and therefore need to be more rigorously tested during the interview process.

[28] One might object that, since the female applicant is given credibility in excess of what is epistemically warranted, she has not suffered a credibility deficit. But, according to Fricker, "credibility deficit" refers to credulity that is less-than-it-would-be-in-the-absence-of-prejudice (Fricker 2007, 17–18). It does not refer solely to credulity that is less than warranted. And I think Fricker's counterfactual test is entirely appropriate given the kind of prejudicial incredulity that warrants our attention.

No candidate is attempting to transmit knowledge with their answers, thus the question of epistemic exclusion does not arise.[29] [30]

Fourth,

> *Baseball Batter:* Suppose a woman grabs a baseball bat with the intent to take some practice swings. A nearby man assumes that, because she is a woman, she does not know how to hold the bat. So he begins telling her how to hold a bat.[31]

This is a testimonial injustice, but there was no attempt (by the victim) to transmit knowledge.

Fifth,

> *Self-Doubt*: Suppose a woman doubts that she herself knows this or that masculine-coded item of knowledge because she is a woman. (I.e. a self-inflicted testimonial injustice.)

This example fits all the required features of testimonial injustice (albeit the hearer and speaker happen to be identical). But nobody has *attempted* to contribute knowledge to a shared pool of understanding and thus nobody has been *thwarted* in such an attempt.

Sixth,

> *Incorrect Yet Reliable Speaker:* Suppose a hearer doubts the testimony of a reliable speaker due to an identity-prejudicial stereotype. Although the speaker is reliable *in general*, in this particular instance the speaker happens to be incorrect.[32]

---

[29] Of course, the interviewees are trying to convince the interviewer of something (namely, that they are qualified), but the answers to the questions themselves are already known by all parties to already be in the interviewer's pool of knowledge. We can also stipulate that the interviewees do not themselves know whether they are the best applicant.

[30] Katherine Hawley reports on some real-world evidence for these kinds of scenarios in connection with testimonial injustice: "In studies involving simulated job applications…women as compared with men, and black people as compared with white, 'must work harder to prove that their performance is ability-based'" (Hawley 2011, 294).

[31] Thanks to Julia Markovits for this example. Julia Markovits's original example was actually more revolting: the man approaches from behind and wraps his arms around the woman, unannounced, to teach her how to hold the bat.

[32] Thanks to David Sobel for suggesting this example.

The speaker has suffered a testimonial injustice, but no knowledge would have been communicated in the absence of the testimonial injustice (since false beliefs do not count as knowledge) and thus testimonial injustice occurred without having thwarted a contribution of knowledge. (Though perhaps we would want to say here that the attempt to communicate knowledge was thwarted in multiple ways – i.e. its thwartation was overdetermined.)

To recap, I have proposed six counterexamples to premise 1's claim that thwarting the submission of knowledge to a shared pool – "epistemic exclusion" – is a necessary feature of testimonial injustice. In *Loan Officer* and *Persuasive Speaker*, knowledge *is* successfully communicated, *even though* the communicator suffered a testimonial injustice. In *Proofreader Interview*, *Baseball Batter*, and *Self-Doubt*, testimonial injustice *did not* thwart the submission of knowledge because no communication of knowledge would have occurred in the absence of the testimonial injustice. The lesson from these counterexamples is that prejudicially doubting someone thwarts knowledge communication *only sometimes*; there are many cases where it does not. The upshot of this lesson is that even if thwarting someone from communicating knowledge were a way of directly harming a person, it could not be *the* direct harm of testimonial injustice.

In response to these counterexamples, Fricker could weaken Premise 1 as follows: 'Testimonial injustice *sometimes* thwarts one's attempt to contribute knowledge.' And the conclusion would then of course be weakened to read: 'Therefore (from 6, 7) testimonial injustice *sometimes* causes a direct harm (because it *sometimes* is a type of Function-Dehumanization).' But the counterexamples I presented here do seem, intuitively, to involve direct harm; the suggested modification would only be relevant if the counterexamples were ones that did not seem to involve any direct harm whatsoever.

Some of my counterexamples do not involve testimony, which may be surprising. Is testimonial injustice without testimony not a contradiction in terms? But even Fricker does not think testimonial injustice requires testimony; she writes: "The phenomenon I call testimonial injustice

is not in fact confined to testimonial exchange, even allowing that we intend testimony in its broadest sense to include all cases of telling" (Fricker 2007, 60). Part of the confusion is simply the chosen label "testimonial injustice." This label is a technical term that refers to prejudicial *credibility deficits*. Credibility assessments often occur in episodes involving testimony, but they also occur in episodes not involving testimony. In an interview, Fricker explains her choice of phrasing: "Sometimes I wonder if I ought to have called it plain 'credibility injustice' or 'assertion injustice' or something. But I had methodological reasons in the book for focusing on testimony, reasons that are connected with a view of Edward Craig's I explore there which puts testimony at the heart of what it is to know" (Dieleman 2012). (I wish she had called it "credibility injustice" (or perhaps "incredulity injustice") so as to avoid this confusion.)

Section 5 - Conclusion

I have explained why testimonial injustice is inherently wrong. The explanation is as follows: Testimonial injustice is a type of negatively valenced, prejudicial incredulity (by definition). When an agent is incredulous regarding another person's sincerity or expertise (in a domain of knowledge where insincerity or non-expertise is negatively valenced by the agent), they thereby disesteem that subject. Being disesteemed by others is a direct harm. Thus the kind of incredulity involved in testimonial injustice is directly harmful. When the distribution of a harm is influenced by identity-prejudice, it is unjust. Thus testimonial injustice is inherently an unjust distribution of harms.

Additional support for my theory comes from its ability to handle counterexamples to rival theories. Those counterexamples were ones in which no capacity to reason was undermined or no attempt to

transmit knowledge was thwarted (either because no attempt was made in the first place or because the credibility deficit was outweighed).

I will review two of those counterexamples here to explain how my theory handles them. First, *Loan Officer* (in which the loan officer gives the male and female applicants too much credibility yet would have given the woman even more credibility in the absence of his sexist beliefs about the competence of businesswomen). In this case, the amount of esteem that the loan officer has for each applicant is comparatively unfair because the difference in goods they each receive is not caused by any difference in what they merit. It is unjust because the difference is caused by prejudice. Although the woman is not harmed in an absolute sense (since the loan officer *does* esteem her competence highly), she is harmed in a comparative sense; in particular, she does not receive *as much* esteem as she would have in the absence of the loan officer's prejudicial stereotype.[33] Second, *Self-Doubt* (in which a woman commits a testimonial injustice against herself). In this case the woman's self-esteem is lower not as a result of the woman's merits but as a result of the woman's sexist prejudice. Thus, she does herself an injustice by distributing a direct harm to herself due to her prejudice.

[33] Compare, for example, an unjustly distributed government handout, in which poor racial minorities receive about half as much money as poor racial non-minorities (all else equal). The racial minorities are not *absolutely* harmed by the handout, but they do benefit less (which is sufficient for distributive injustice).

Bibliography

Alcoff, Linda Martin. 1991. “The Problem of Speaking for Others.” *Cultural Critique*, no. 20: 5–32. https://doi.org/10.2307/1354221.

———. 2010. “Epistemic Identities.” *Episteme* 7 (2): 128–37. https://doi.org/10.3366/E1742360010000869.

Anderson, Elizabeth. 2010. “The Fundamental Disagreement between Luck Egalitarians and Relational Egalitarians1.” *Canadian Journal of Philosophy; Edmonton* 40: 1–23.

———. 2012. “Epistemic Justice as a Virtue of Social Institutions.” *Social Epistemology* 26 (2): 163–73. https://doi.org/10.1080/02691728.2011.652211.

Aristotle. 1941. *The Basic Works of Aristotle*. Translated by Richard McKeon. New York: Random House.

Battaly, Heather. 2017. “Testimonial Injustice, Epistemic Vice, and Vice Epistemology.” In *The Routledge Handbook of Epistemic Injustice*, edited by Ian James Kidd, Jose Medina, and Gaile Pohlhaus, Jr., 1st ed., 456. London: Routledge. https://doi.org/10.4324/9781315212043-22.

Begby, Endre. 2013. “The Epistemology of Prejudice.” *Thought: A Journal of Philosophy* 2 (2): 90–99. https://doi.org/10.1002/tht3.71.

Bruin, Boudewijn de. 2014. “Self-Fulfilling Epistemic Injustice.” SSRN Scholarly Paper ID 2588430. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2588430.

Congdon, Matthew. 2017. “What’s Wrong with Epistemic Injustice?” In *The Routledge Handbook of Epistemic Injustice*, edited by Ian James Kidd, Jose Medina, and Gaile Pohlhaus, Jr., 1st ed., 456. London: Routledge. https://doi.org/10.4324/9781315212043-22.

Davis, Emmalon. 2016. “Typecasts, Tokens, and Spokespersons: A Case for Credibility Excess as Testimonial Injustice.” *Hypatia* 31 (3): 485–501. https://doi.org/10.1111/hypa.12251.

Deardorff, Julie. 2000. “Shame Returns as a Punishment.” *Chicagotribune.Com*, April 12, 2000. https://www.chicagotribune.com/news/ct-xpm-2000-04-12-0004120235-story.html.

Dieleman, Susan. 2012. “An Interview with Miranda Fricker.” *Social Epistemology* 26 (2): 253–61. https://doi.org/10.1080/02691728.2011.652216.

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

———. 2015. “Epistemic Contribution as a Central Human Capability.” In *The Equal Society: Essays on Equality in Theory and Practice*, edited by George Hull. Lanham, Maryland: Lexington Books.

———. 2016. “Fault and No-Fault Responsibility for Implicit Prejudice—A Space for Epistemic ‘Agent-Regret.’” In *The Epistemic Life of Groups: Essays in the Epistemology of Collectives*, edited by Michael Brady and Miranda. Fricker, First edition. Oxford, United Kingdom: Oxford University Press.

Gardner, Molly. Forthcoming. “What’s the Harm?”

Hawley, Katherine. 2011. “Knowing How and Epistemic Injustice.” In *Knowing How: Essays on Knowledge, Mind, and Action*, edited by John Bengson and Marc A. Moffett, 283–99. Oxford University Press.

Helfand, Zach. 2018. “What It Took to Write About Baseball as a Woman,” October 15, 2018. https://www.newyorker.com/magazine/2018/10/22/what-it-took-to-write-about-baseball-as-a-woman.

Hunt, Elle. 2017. “Alien Intelligence: The Extraordinary Minds of Octopuses and Other Cephalopods.” *The Guardian*, March 28, 2017, sec. Environment. https://www.theguardian.com/environment/2017/mar/28/alien-intelligence-the-extraordinary-minds-of-octopuses-and-other-cephalopods.

Jones, Karen. 2001. “The Politics of Credibility.” In *A Mind of One’s Own: Feminist Essays on Reason and Objectivity*, edited by Louise M. Antony and Charlotte Witt, 2nd ed. Boulder, Colo.: Westview Press.

Lee, Harper. 2002. *To Kill a Mockingbird*. First Perennial classics edition. New York: HarperCollins. http://cornell.lib.overdrive.com/ContentDetails.htm?ID=7734E6F7-90B5-4940-8CE4-5FE2D80DB821.

Maitra, Ishani. 2010. “The Nature of Epistemic Injustice.” *Philosophical Books* 51 (4): 195–211.

Manne, Kate. 2016. “Humanism: A Critique.” *Social Theory & Practice* 42 (2): 389–415.

McKinnon, Rachel. 2016. “Epistemic Injustice.” *Philosophy Compass* 11 (8): 437–46. https://doi.org/10.1111/phc3.12336.

Medina, José. 2011. “The Relevance of Credibility Excess in a Proportional View of Epistemic Injustice: Differential Epistemic Authority and the Social Imaginary.” *Social Epistemology* 25 (1): 15–35. https://doi.org/10.1080/02691728.2010.534568.

Origgi, Gloria, and Serena Ciranna. 2017. “Epistemic Injustice: The Case of Digital Environments.” In *The Routledge Handbook of Epistemic Injustice*, edited by Ian James

Kidd, Jose Medina, and Gaile Pohlhaus, Jr., 1st ed., 456. London: Routledge. https://doi.org/10.4324/9781315212043-22.

Pohlhaus Jr., Gaile. 2012. “Relational Knowing and Epistemic Injustice: Toward a Theory of ‘Willful Hermeneutical Ignorance.’” *Hypatia* 27 (4): 715–35.

———. 2014. “Discerning the Primary Epistemic Harm in Cases of Testimonial Injustice.” *Social Epistemology* 28 (2): 99–114. https://doi.org/10.1080/02691728.2013.782581.

Riggs, Wayne. 2012. “Culpability for Epistemic Injustice: Deontic or Aretetic?” *Social Epistemology* 26 (2): 149–62.

Rosenbaum, Traci. 2019. “2 Men Who Lied about Military Service Must Wear Sign That Says ‘I Am a Liar,’ Judge Rules.” *USA TODAY*, August 27, 2019. https://www.usatoday.com/story/news/nation/2019/08/27/two-montana-men-who-lied-being-veterans-sentenced/2128167001/.

Sen, Amartya. 1985. “Well-Being, Agency and Freedom: The Dewey Lectures 1984.” *The Journal of Philosophy* 82 (4): 169–221. https://doi.org/10.2307/2026184.

Singer, Peter. 2009. “Speciesism and Moral Status.” *Metaphilosophy* 40 (3–4): 567–81. https://doi.org/10.1111/j.1467-9973.2009.01608.x.

Sullivan, Shannon. 2017. “On the Harms of Epistemic Injustice : Pragmatism and Transactional Epistemology.” In *The Routledge Handbook of Epistemic Injustice*, edited by Ian James Kidd, Jose Medina, and Gaile Pohlhaus, Jr., 1st ed., 456. London: Routledge. https://doi.org/10.4324/9781315212043-22.

Tanesini, Alessandra. 2018. “Caring for Esteem and Intellectual Reputation: Some Epistemic Benefits and Harms.” *Royal Institute of Philosophy Supplement* 84: 47–67. https://doi.org/10.1017/s1358246118000541.

Wanderer, Jeremy. 2012. “Addressing Testimonial Injustice: Being Ignored and Being Rejected.” *Philosophical Quarterly* 62 (246): 148–69.

Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert. 2013. “Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas.” *Behavior Research Methods* 45 (4): 1191–1207. https://doi.org/10.3758/s13428-012-0314-x.

Williams, Bernard. 1981. *Moral Luck: Philosophical Papers, 1973-1980*. Cambridge [Cambridgeshire]: Cambridge University Press.

Wilson, Judith. 1991. “Down to The Crossroads: The Art of Alison Saar.” *Callaloo* 14 (1): 107–23. https://doi.org/10.2307/2931444.

Hypocrisy as Selfish Self-Exceptionalism

Section 1 - Introduction

What is hypocrisy? Why is it bad? How bad is it?

I propose that hypocrisy is selfish self-exceptionalism, which is when a person holds others to a more demanding standard than that to which she holds herself due to a selfish bias. This theory explains what hypocrisy is. I also argue that the intrinsic selfishness of hypocrisy, in the context of committing to normative standards, explains why hypocrisy is bad.

That *selfishness* explains why hypocrisy is bad (and is a necessary condition of hypocrisy) is omitted or underemphasized in the philosophical literature. The selfishness of hypocrisy pops out at you when considering contrasting, non-hypocritical cases in which an agent holds others to a *lower* standard than that to which she holds herself. I call these (non-hypocritical) kinds of cases self*less* self-exceptionalist cases. Failure to pay attention to these cases leaves hypocrisy's theorists unable to differentiate between hypocritical and non-hypocritical double standards. That is, hypocrisy as selfish self-exceptionalism clarifies the continuities and discontinuities between hypocritical and non-hypocritical double-standards.

The following case exemplifies hypocrisy as selfish self-exceptionalism

> *The Demanding Professor*: Harvey, a professor, holds others to a high standard of research productivity. However, he does not hold himself to as high a standard.

In contrast, the following case exemplifies self*less* self-exceptionalism (which is *not* hypocritical):

> *The Productive Professor*: Cathy, a professor, holds herself to a high standard of research productivity. However, she does not hold others to as high a standard.

In both cases, the professor is guilty of a double standard: the professor has one standard for some and a different standard for others even though there are no impartial grounds for these distinct standards. But, intuitively, only Harvey's attitudes are hypocritical (and offensive).

In Section 2, I explain the theory of hypocrisy as selfish self-exceptionalism. In section 3, I apply the theory of selfish self-exceptionalism to the politically salient double standard of sexist conversational norms. In Section 4, I situate my theory in the existing literature on hypocrisy. In Section 5, I respond to objections.

Section 2 – Hypocrisy Defined and Evaluated

In this section I will explain and evaluate my theory of hypocrisy by working through a taxonomy of sub-categories: Double-Standards → Exceptionalism → Self-Exceptionalism → Selfish Self-Exceptionalism.

To hold a "double-standard" is to arbitrarily hold one group to one standard and at least one other group to a different standard. More concisely, it is to hold arbitrarily distinct standards for different groups. For example,

> *Dolls and Trucks:* Girls ought to prefer dolls, boys ought to prefer trucks.[34]

This is a double standard since gender-differences are not objectively relevant to what type of toy one ought to prefer.

The standards involved in a double-standard will either be equally (un)demanding or one will be more demanding than the other.[35] When a standard is harder to comply with than another, it is "higher." For now I will assume we have an intuitive grasp of the notion of holding people to different standards. After I have finished defining hypocrisy, I will explore this notion in more depth.

Next we have "exceptionalism." I repurpose this word from the phrase "American Exceptionalism," which is the ideology that the USA is somehow special, i.e. that rules that apply to countries in general do not apply to the USA - that the USA ought follow different norms than other countries.

For my purposes, "exceptionalism" means to hold a double-standard *out of partiality*. For example,

> *Teacher's Pet:* A teacher is strict when her students break the rules, in general, but is lenient when her favorite student breaks the rules.

The teacher holds one person (her favorite student) to a different standard than another person or group (the other students) and does so out of partiality towards that student. (We could imagine a contrasting scenario in which the teacher holds the student to an easier standard *not* out of partiality, e.g. because the student is so well-behaved in general that he deserves leniency or e.g. because the student has a behavioral disability, etc.). A reason is partial if it is grounded in a special attitude the agent holds towards one of the subjects – it is a form of bias. To test for partiality, we can imagine an impartial observer, limited to the same information as the original agent, and ask if they would make the same distinction in standards as has the original agent. If the impartial observer would *not* apply the same distinction in standards to the subjects as did the original agent, then said agent must have been partially biased (at least to some degree).

In contrast to *Teacher's Pet*, the double standard involved in *Dolls and Trucks* is not partially biased, which is why it falls *outside* the category of exceptionalism. (As evidence that the double standard in

[34] Thanks to Julia Markovits for this suggesting this example.
[35] In theory, the demandingness of two standards could also be incomparable.

Dolls and Trucks is not partially motivated, consider that boys are equally as likely as girls to believe in the double standard.)

Exceptionalism need not be immoral. For example,

> *Biased Parent*: A father forgives his daughter for a crime she committed, though he is not disposed to forgive anyone else for the same kind of crime.

The parental bias is paradigmatically partial and so the father is engaging in exceptionalism. But I do not think his uneven dispositions strike us as morally offensive. (Let us stipulate that there are no significantly bad consequences to his having these uneven forgiveness-dispositions.)

Moving on now to *self*-exceptionalism, which refers to those cases of exceptionalism in which one of the two groups of people is the agent herself. Consider, for example,

> *Bizarre Dietician*: An amateur dietician prescribes red apples for others, but green apples for herself. (The dietician likes the taste of each kind of apple equally.)

Here the agent has one norm ("One ought to eat X") for others and a different norm ("One ought to eat Y") for herself. It is hard to imagine what kind of delusion explains the dietician's bias, but I will stipulate that it is partial. Note that each standard is equally easy to comply with.

It is common with distinct standards in the same category, for one standard to be more demanding, or "higher," than the other. When one holds others to a *lower* standard than one holds others, one's self-exceptionalism is typically self*less*, as in the following (repeated from the introduction):

> *The Productive Professor*: Cathy, a professor, holds herself to an incredibly high standard of research productivity. However, she does not hold others to as high a standard.

Finally we arrive at hypocrisy, with cases of selfish self-exceptionalism. Such cases involve holding others to a *higher* standard than oneself. Here are a couple of examples:

> *Hypocritical Driver:* An automobile driver believes that others are acting very badly when they speed, but believes that she herself is acting only mildly badly when she speeds.

> *Hypocritical Carpenter:* A carpenter criticizes her brother for having a splinter in his eye but does not criticize (or bother to remove) a beam of wood from her own eye.[36]

In each of these examples, the agent holds others to higher standards than that to which they hold themselves. And in both examples, the best explanation for why the agent holds the double standard in question is that the agent has a selfish bias.

Is there a difference between selfish self-exceptionalism and "holding others to a higher standard than one holds oneself"? And if so, which of the two best captures the extension of "hypocrisy"? There is a difference, as the following example illustrates:

> *Daredevil Circus*: The head of auditions to the circus's daredevil performers troupe is herself a daredevil. Of the applicants she rejects, many are moderately better at stunts than she. Thus she holds others to a more demanding standard than she does herself. However, she does this because it is a dangerous job and she is less willing to risk the lives of others than to risk her own life. This explains why she demands *much* more excellence in the abilities of other daredevils than in her own.

[36] Although unrealistic (is the carpenter a giant or something?) it is historically and etymologically important. Etymologically, use of the word "hypocrite" can be traced back to the New Testament. There, the protagonist accuses others of a variety of transgressions, frequently using the word "hypocrite" to mean "wrongdoer" in a very generic sense. Of relevance is the following passage:

> "Judge not, that ye be not judged. For with what judgment ye judge, ye shall be judged; and with what measure ye mete, it shall be measured to you again. And why beholdest thou the mote that is in thy brother's eye, but considerest not the beam that is in thine own eye? Or how wilt thou say to thy brother, 'Let me pull out the mote out of thine eye,' and behold, a beam is in thine own eye? Thou *hypocrite,* first cast out the beam out of thine own eye, and then shalt thou see clearly to cast out the mote out of thy brother's eye." Matthew 7:1-5 KJ21 (emphasis added)

A similar exchange is found in the Talmud:

> "And it came to pass in the days when the judges [were] judged…If, e.g., the judge said to them: 'Takeout the toothpick from thy tooth,' they answered: 'If thou wilt take the beam out of thy eyes, I will remove the toothpick.' (i.e., if the judge accused one of a small transgression, the accused said to him: 'Thou thyself art a greater sinner than I am')." ("Tractate Bava Batra: Chapter 1" n.d.)

By my intuitions, the head of daredevil auditions is not hypocritical. The explanation for why she is not is that "hypocrite" refers only to those instances of "holding others to a higher standard" that are selfishly biased.

I will clarify the concept of holding others to a higher standard by describing some of the ways in which it occurs:

1. You could think that both yourself and another should not violate norm X, but also think that when you violate X it is not as bad as when some other person does (explored as a form of hypocrisy by Valdesolo and DeSteno 2007);
2. You could think that a more demanding version of the same sort of norm applies to another person than applies to yourself (e.g. they should give 20% of their income to charity, but I only need to give 10% of my income to charity);
3. You could be disposed to a harsher reaction towards another person when they violate norm X than you do towards yourself when you violate X (e.g. you blame, punish or forgive another for violating X, but you do not blame, punish or forgive yourself (as much) for violating X)
4. You could be more readily disposed to believe that you occupy a non-arbitrary exception to the norm than you are to allow for such exceptions on behalf of others.
    a. To be more "readily disposed" itself has at two different forms: 1) you could be extra-motivated or eager to find reasons supporting your being a legitimate exception to the norm; or 2) you could make the "fundamental attribution error" which occurs when one emphasizes situational factors to explain one's own mistakes but character traits to explain the mistakes of others.[37]

One further clarification is needed before concluding this descriptive definition. The "self" in "self-exceptionalism" must include team-identification. For example:

> *Bush's Minimum Sentencing:* "I respect the jury's verdict. But I have concluded that the prison sentence given to Mr. Libby is excessive." With these words former President

[37] For example, welfare recipients who resent the laziness of other welfare recipients but think that there are legitimate mitigating circumstances in their own case.

> George W. Bush commuted the sentence of I. Lewis "Scooter" Libby, Jr., for obstruction of justice and leaking the identity of CIA operative Valerie Plame. Critics of the decision noted that Libby actually had received the minimum sentence allowable for his offense under the law, and that many of Libby's supporters, including the Bush administration, were actively pressing for mandatory minimum sentencing laws at a national level (Goodman, 2007). Accordingly, critics of the decision saw it as a textbook case of moral hypocrisy: different rules were being applied to Bush's underling, Libby, than to everyone else in the United States."[38]

Contrast Bush's hypocritical action in the case above with a hypothetical case in which Bush pardons someone to whom he has no team-identification (perhaps he just likes the person's style, was in a particularly good mood when the case was presented to him, etc.). That latter, hypothetical action would also involve a double standard (probably a wrongful one) but would not be self-exceptional and so would not be hypocritical.

The final version of my definition is as follows:

> **Hypocrisy as Selfish Self-Exceptionalism:** A person's attitudes or behavior are hypocritical iff she holds others (or out-group entities) to a more demanding standard than that to which she holds herself (or in-group entities) in a case where all else is equal, and does so due to a selfish bias.

Having defined hypocrisy, the next question to ask is why it is bad. My thesis is that hypocrisy is bad because it is selfish and that the badness of selfishness varies depending on the context and degree of selfishness. In some contexts, hypocrisy will only be mildly bad, such as:

> *Fashion Hypocrisy:* Bill sneers at those who wear socks and sandals, but has no compunction wearing socks with sandals himself.

[38] (Rai and Holyoak 2014, 2475) Includes citation of (Goodman 2007)

Bill's hypocrisy is only mildly bad. Hypocrisy will tend to be more serious when the norms in question are moral and, not only moral, but with significantly important consequences. Not just the consequences of a particular instance of hypocrisy, but also with other dispositions we would predict of the agent. For example, part of what makes Bill's hypocrisy relatively benign is not only that his fashion attitudes towards socks and sandals is inconsequential, but also that we would not anticipate, on the basis of the character revealed by this particular selfish self-exceptionalism to a more serious instance of hypocrisy. In contrast, in *Bush's Minimal Sentencing*, we *are* anticipating that the agent's bias will repeat itself in additional, serious circumstances.

That, I believe, explains the badness of hypocrisy, though there is more to say about why hypocrisy is considered *so* distasteful in our society. Sure, the hypocrite is selfish…but selfishness is such an everyday vice. Why is hypocrisy considered *so* taboo in our society? Part of the explanation, I think, is that the norm "do not be a hypocrite" is so broadly agreed upon. Criticizing a person for violating a norm to which that person does not subscribe is unsatisfying (unless one is, say, communicating solely within one's political or cultural bubble about an outsider). A norm that *everyone* agrees with will therefore be invoked disproportionately frequently with respect to the seriousness of violating the norm. (In contrast, a more serious norm about which there was widespread *disagreement* would be invoked relatively infrequently.)

Section 3 - Double Standards

Sometimes the context in which hypocrisy occurs is very serious. I touched briefly on the relationship between hypocritical and non-hypocritical double standards when I discussed the contrast between self- and non-self-exceptionalism (*Teacher's Pet*). In this section I analyze another politically salient case of double standards, namely sexist conversational norms, through the lens of selfish self-exceptionalism. My conclusion will be that men who hold sexist conversational norms *are* hypocrites, though only in general and mostly just the adult men.

By "sexist conversational norms," I have in mind norms governing such behavior as interrupting others, talking over others, speaking with an angry affect, etc., all of which are often treated as more acceptable when performed by a man than by a woman. I take for granted that there is no legitimate, objective justification for this double standard.

For simplicity, I will restrict my analysis to two types of agents: a cis-man and a cis-woman. Let us start with the woman. She believes (at least implicitly) that it is generally more acceptable when men

interrupt women, speak with angry effect, etc. then when women do. She therefore holds men to a less demanding standard then herself (and other women) and is not hypocritical.

Turn now to the man who holds these same beliefs. This man holds roughly half the population to a more demanding standard than he does himself. Another description of the same case is that he holds an out-group (women) to a more demanding standard than he does his in-group (men). So on both the individualist and team-based definition, he has at least one of the markers of hypocrisy. The next question to ask is whether the man has the other marker: is he selfishly motivated?

I think the right answer to that question will vary on a case by case basis. Some men hold sexist conversational norms innocently and others do not. The plausibility of the former kind of men is controversial, so I will elaborate. Consider again the woman who holds such norms. Why does she do so? Presumably because of the social environment in which she was raised. This environment teaches young girls these norms in at least a couple of different ways: 1) By example (i.e. they witness men getting away with relatively impolite conversational behavior but do not witness women getting away with the same kind of behavior) and 2) by direct discipline (i.e. they are chastised relatively more severely than boys for relatively impolite conversational behavior). At some point (hopefully) the woman will be in a position to realize that the gendered differences in the acceptability of talking over others are arbitrary and will change her attitudes towards that behavior. Now let us turn back to the sexist man. Presumably the man who holds this double standard was subject to similar lessons as a child. He witnessed the same behavior amongst adults and was chastised relatively less severely than girls for relatively impolite conversational behavior. After considering the parallel structure of the woman's and man's upbringing, it is not obvious to me that the man must be selfishly motivated in order to hold sexist conversational norms. This explains why I think that whether it is hypocritical for a man to hold these norms will vary on a case by case basis.

But by adulthood, in today's world, there is enough information, and sufficiently widespread, for everyone to be in a position to recognize the conversational double standard. What would motivate an adult man to resist evidence that conversational norms are gendered and arbitrarily so? Those double standards work in the man's favor: they excuse his relatively less-controlled behavior. So its plausible that resistance to recognizing this double standard for what it is is selfishly motivated and therefore hypocritical.

Section 4 - Situating Hypocrisy as Selfish Self-Exceptionalism in the Literature

There are two distinct concepts referred to by "hypocrisy" in the philosophical literature; I will call these concepts *Hypocrisy as Moral Pretense* and *Hypocrisy as Differential Blaming Disposition.* The former (moral pretense) is not related to selfish self-exceptionalism in any philosophically interesting way. In fact, I will discuss some evidence that it may reflect a fading linguistic usage. The latter (differential blaming disposition) is a subcategory of selfish self-exceptionalism.[39] So in one sense, differential blaming disposition is not in competition with my view – it is just narrower in scope. But the extant

[39] In particular, a subcategory of the third type of "holding others to a higher standard" – see the end of section 2

evaluative analyses of differential blaming dispositions do not emphasize the selfishness of hypocrisy, which is a mistake.

Section 4.1 - Hypocrisy as Moral Pretense

Hypocrisy as moral pretense is widely endorsed, e.g. in (Ryle 1984, 173), (Szabados 1979), (Kittay 1982), (McKinnon 1991), (Soifer and Szabados 1998), (Szabados and Soifer 2004), (Wallace 2010), (Rossi 2018), and (https://www.merriam-webster.com/dictionary/hypocrisy, 2018). Moral pretense means deceiving (or, at least, attempting to deceive) others into thinking one is more moral than one actually is. The following example illustrates:

> *Sexist Employer*: "a male employer who, although an unregenerate sexist but desiring to impress some woman with his 'open mindedness', hires a well-qualified woman for what is traditionally a man's position." (Kittay 1982, 277)

In this example, the employer attempts to deceive other women into thinking he is more moral than he really is. (The relevant perspective, must, of course, be those the employer is deceiving, given that, from the perspective of an unregenerate sexist, acting 'open-minded' is *less* moral than acting 'close-minded' (though the scare-quotes are necessary only from the perspective of the employer)).

In my social milieu, "hypocrisy" does not mean moral pretense. In fact, the first time I encountered that usage was in academic philosophy essays. Several other philosophers of hypocrisy also do not take moral pretense to be the (only) definition, and some of them do not mention that view at all. Thus, I will distinguish between two dissenting positions one might take to the Hypocrisy as Moral Pretense view. First, a Moderate Dissenting View, according to which moral pretense is one use of the word hypocrisy (i.e. moral pretense, of the right kind, is a sufficient, but unnecessary condition for hypocrisy) (Crisp and Cowton 1994) (Wallace 2010) (Alicke, Gordon, and Rose 2012). Second, a Radical Dissenting View, according to which moral pretense is *not* hypocrisy (i.e. moral pretense is *neither* a sufficient nor a necessary condition for hypocrisy) (Fritz and Miller 2018) (Roadevin 2018).

Given the large number of essays on hypocrisy as moral pretense, the radical dissenting view seems prima facie to be radically implausible. One possibility, however, is that the folk meaning of hypocrisy used to be moral pretense and has been transitioning into meaning selfish self-exceptionalism. That is, the radical dissenting view is not *yet* true across the board but is true in many linguistic communities and is on the path to becoming true across the board. In any case, the selfish self-exceptionalism thesis is of philosophical interest even if merely the moderate dissenting view is correct. For in that case my thesis would simply be about one widespread usage of hypocrisy. Still, the debate between the radical and moderate dissenting views here is of some interest in itself and has *some* degree of impact on how philosophically interesting the selfish self-exceptionalist ought to be. So I will spend the next couple of pages arguing in favor of the radical dissenting view.

My arguments for the claim that moral pretense (of the right sort) is a kind of hypocrisy only to a *minority* of people rely on an experimental philosophy study by Alicke, Gordon, and Rose (Alicke, Gordon, and Rose 2012). Alicke et al. surveyed hundreds of undergraduate students, asking them to judge whether hypocritical behavior was on display in a variety of hypothetical scenarios. I will review a couple of the scenarios. Besides noting that moral pretense usage appears to be held by a minority of

respondents I will also take the opportunity to apply the selfish self-exceptionalism theory. Each scenario has three versions; each respondent was randomly presented with *one* version from each scenario. The scenarios I review here progress as follows from one scenario-version to the next: honest and forthcoming agent → neither deceptive nor forthcoming agent → lying agent.

> *Reformed Parent version 1:* "A parent tells his 17 year-old son that although he drank alcohol and smoked pot when he was 17 years old, he does not want his son to do so." (Alicke, Gordon, and Rose 2012, 691)

57% of respondents labeled the parent hypocritical. In version 2, the parent does not tell his son that he used to drink and smoke; 54% labeled the parent hypocritical. In version 3, the parent outright lies, telling his son that he never drank or smoke; 69% labeled the parent hypocritical. The big question for moral pretense theorists is why 57% of respondents labeled the version 1 parent a hypocrite given that version 1 does not involve any moral pretense.

Hypocrisy as selfish self-exceptionalism can explain these data points. The scenario as written is ambiguous with regards to whether the parent is a self-exceptionalist. If the parent *now* believes that *his* past behavior was *in*excusable, then he is *now* holding his past self to the same standards as that to which he holds his son. If, however, he *now* believes his past, teenage behavior *was* excusable, then he is holding his son to a more demanding standard. Given this ambiguity, it is unsurprising that approximately half of respondents labeled the parent a hypocrite and half did not.

Version one of the second scenario goes as follows:

> *Premarital Sex version 1:* "Jane believes it is wrong to have premarital sex and shares her attitudes with other people. Jane had pre-marital sex and tells this to others when sharing her attitudes." (Alicke, Gordon, and Rose 2012, 691)

73% of respondents labeled Jane a hypocrite in this version. In version 2, Jane does not mention to others that she had pre-marital sex; 94% labeled her a hypocrite. In version 3, Jane tells everyone she was a virgin when she was married; 96% labeled her a hypocrite. Again, moral-pretense-theorists will have trouble explaining why 73% of respondents thought that Jane in version 1 was a hypocrite. The selfish self-exceptionalism theory can explain the responses. In version 1, it is ambiguous whether Jane has reformed her behavior and evaluative beliefs regarding pre-marital sex. If she used to believe pre-marital sex were permissible, back when she was engaging in it, but no longer believes so (and is no longer engaging in premarital sex) then she was not making an exception of herself at any single point in time. Given the topic, it seems a bit more likely that Jane has not changed her mind. This is just a guess based on anecdotal evidence of people who are against pre-marital sex: such people seem to always have had those beliefs. That kind of evidence is not particularly rigorous but it does give us insight into what the survey respondents might have been thinking. (In *Reformed Parent*, in contrast, it does seem quite plausible that one would come to regret drinking and smoking as a teenager).

Section 4.2 - Hypocrisy as Differential Blaming Disposition

Jay Wallace's example and subsequent evaluation exemplifies the hypocrisy as differential blaming disposition approach[40]:

> "Suppose I blame you for your dishonesty when I have regularly been dishonest in my interactions with you, and suppose I also fail to reflect on and come to terms with my dishonest behavior in the past. [This] shows that I take your interests to be less important than my own, and that I ascribe to myself a moral standing that I am not willing to grant to you. We all have an interest in being protected from the kind of social disapproval and opprobrium that are involved in blame...This offends against a presumption in favor of the equal standing of persons that I take to be fundamental to moral thought." (Wallace 2010, 328)

Recall that, according to my Hypocrisy as Selfish Self-Exceptionalism theory, differential blaming dispositions *are* a type of hypocrisy. In the above example, the hypocrite holds someone else to a standard of honesty that is higher than that to which they hold themself and does so for selfish reasons. So in applying the word 'hypocrisy' to this case, there is no disagreement between my view and the Differential Blaming Disposition view.[41]

When we come to the evaluative analysis, based on offense "against a presumption in favor of the equal standing of persons," this is close to my own analysis. But this way of framing things makes it sound like there is no evaluative difference between double standards in general and hypocrisy in particular. As I explained at length in section 2, there are a number of characteristics of hypocrisy that make it a narrower phenomenon than double standards (though there are also a number of ways in which it is a broader concept that differential blaming dispositions).[42]

---

[40] Though the phrase "Differential Blaming Disposition" was coined in (Fritz and Miller 2018).

[41] Many Differential Blaming Disposition theorists are particularly interested in why being a hypocrite undermines one's standing to blame. So it is unsurprising that those theorists are focusing on a relatively narrower scope of hypocrisy.

[42] Wallace does briefly consider the "higher standards" understanding of hypocrisy:

> "One might say that hypocrites hold other people to higher standards than they hold themselves to (as several commentators on earlier versions of this article suggested to me). This way of speaking can be misleading, however, since not all applications of double standards offend against the presumption of equal standing that I have identified. In the paradigm cases I am trying to analyze, hypocrites apply double standards precisely by accepting a threshold for subjecting others to opprobrium that is lower than the threshold they apply to their own case. Double standards that do not involve this kind of differential treatment would not necessarily attract the moral objection that I am endeavoring to locate. (Consider, e.g., television commentators who deploy labile criteria when reaching judgments about the quality of play of different World Cup teams.)" (Wallace 2010, 333)

That is, Wallace thinks that holding others to higher standards is a broader category than having a differential blaming disposition and therefore is not the concept he is trying to evaluate. In other words, Wallace takes the topic to *only* be blame, and thus considers non-blame types of standard-holding to be irrelevant.

Another reason to think that a selfish bias is the best explanation for the wrong of differential blaming dispositions is that it provides a defense against Daniela Dover's surprising conclusion that differential blaming dispositions are not bad. Dover's argument focuses on the useful aspects of criticism: getting feedback from others on our behavior, debating our moral transgressions openly, etc. Dover writes:

> "This view [that 'our first obligation is to correct our own failings and not to concern ourselves with the failings of others' (Smith 2007, 480)] seems to me to overlook the extent to which we depend on the criticism of others to figure out what our own failings are in the first place. Of course it is true that we cannot be more obligated to correct the failings of others than to correct our own…But this hardly justifies adopting a regime in which we refrain even from concerning ourselves with the failings of others—or from articulating our concerns to others—because we have not yet made enough progress in correcting our own. Such a regime recommends a degree of moral self-reliance that seems unwarranted, given how frequently others see us more sharply than we see ourselves." (Dover 2019, 394)

That is, Wallace's attempt to ground the wrong of (a subcategory of) hypocrisy in our interest against being blamed is problematic since we often lack such an interest in the first place. Daniela Dover goes on to say "My discussion of these cases may not fully persuade committed defenders of the [anti-hypocritical-criticism] norm: broad social acceptance of that norm has left its mark on our intuitions" (Dover 2019, 391). I do not have strong intuitions either way here, but even if Dover is right that we do not have an all-things-considered interest against being morally criticized, there is still something bad about an agent holding themselves to the wrong standard due to a selfish bias. In general, if the consequences of someone's actions are morally benign, but the reasons for their actions (i.e. the motives or bias that explains their choices) is offensive, then we still have a reason to be offended by the whole action-motive package.

Section 5 - Objections and Replies

Section 5.1 - Tom the Pastor

In this objection we have a case that is intuitively hypocritical yet in which the agent does not selfishly hold others to higher standards. Richard Arneson writes

> *Tom the Pastor:* "Tom is the pastor of a church. He preaches to his congregation that they must avoid adultery, fornication, and other sexual behaviors that offend against church doctrine, which reflects divine command…Tom regularly engages in fornication, adultery, and other sexual behaviors that offend against church doctrine…There is discrepancy between what he preaches and what he practices and also between what he publicly professes and what he privately believes. This is garden variety hypocrisy. His sermons are deceptive. His public condemnations of illicit sex from the pulpit convey to his listeners the message that he himself accepts the norms to which he is demanding they conform..."[43]

Arneson goes on to explain that a variety of different motivations could be stipulated. Tom could be motivated by spite for religious people, for example. *Tom the Pastor* is also important because it relies on a common first-pass understanding of hypocrisy that is relatively absent from the academic literature, namely that hypocrisy *just is* not practicing what one preaches.[44]

This example tests two aspects of my theory. First, Tom publicly pretends to hold others to a higher standard than that to which he privately holds himself though he does not genuinely hold others to a higher standard. Second Tom does so for spiteful reasons rather than selfish ones. If Tom is, indeed, a hypocrite, then something has to give.

First, my intuitions are not clear as to whether publicly pretending to hold others to a higher standard (e.g. by preaching) counts as hypocritical (even if selfishly motivated). Actually, my intuitions are not clear on whether such behavior is non-hypocritical either! I am content to say that if such cases are hypocritical, then *pretending* to hold others to a higher standard should be counted as a type of holding others to a higher standard (full-stop). In this case my analysis of ways in which others could be held to a higher standard would be expanded to note that one can do so with publicly-pretended attitudes *or* genuine attitudes. And if such cases are not hypocritical, then the reason would be because "hypocrisy" describes only cases of genuinely holding others to a higher standard.

Our intuitions may be biased by Tom's profession. Alicke et al. found that priests seemed to be judged to be hypocritical simply for being priests. In their study, survey respondents generally granted that weakness of will was not hypocritical, especially if the weakness of will was a one-time lapse. Only 11% thought that an anti-drug activist was not a hypocrite for taking drugs just once in a moment of weakness. But 83% thought that an anti-adultery-preaching priest was a hypocrite for giving in to a married woman's advances just once in a moment of weakness. (Alicke, Gordon, and Rose 2012, 680–81).

So let us consider a case of pretended standards that does not involve a priest. Consider, for your own intuitions, the following scenario.

---

[43] In comments on a draft of this paper presented at the [Removed for Anonymous Review].
[44] Hurka implicitly endorses this view, albeit in a non-academic paper (Hurka 1994)

Pumping Iron: In the movie *Pumping Iron,* Arnold Schwarzenegger wants to win a bodybuilding competition. He tells a fellow competitor (a foreigner who is unfamiliar with the precise norms of the competition) that one is supposed to scream loudly while flexing on stage (for those unfamiliar with the competition, the screaming part is a violation of the competition's norms). Schwarzenegger's competitor follows the norm Schwarzenegger recommended and, as a result, is disqualified. In this case Schwarzenegger preaches a norm he does not follow. (We could also easily alter the case to have his motivations match Tom the pastor's motivations; i.e. we could stipulate either that Schwarzenegger was selfishly motivated or just spitefully motivated, etc.)[45] Is Scharzenegger a hypocrite?

What about the second problem, that of spitefully motivatations? To make things simpler, let us consider a case in which the agent's attitudes are genuine (and the agent is not a priest):

> *Allegedly Hypocritical Driver:* An automobile driver criticizes other drivers when they speed (e.g. with her middle finger), but does not subject herself to any criticism when she speeds. She does this for spiteful reasons.[46]

Spite is an inherently other-directed attitude, so it makes sense that the driver likes to criticize others for spiteful reasons but has no corresponding motivation to criticize herself. But spite is compatible with selfishness and I find it implausible that the *Allegedly Hypocritical Driver* is not selfishly motivated.

Of course, as Arneson points out elsewhere in his comments, and as I now admit (in Section 4.2), my thesis would still be of philosophical interest if moral pretense and selfish self-exceptionalism were simply different kinds of hypocrisy.

Section 5.2 - Accuracy vs. Consistency of Norm Application

Arneson writes, "If norms are being misapplied, it is better that their application more closely approximates correct application. So, better that the standards be inconsistently applied rather than consistently badly applied.

Consider EASY. Easy applies very relaxed, undemanding standards to himself, and consistently applies these same standards to everybody else. Very few homicides are wrongful homicides, according to Easy; very few thefts are wrongful thefts; very few rapes are wrongful rapes, and so on. Easy's standards are much too relaxed. If so, then it is better that Easy apply more demanding standards, closer to the truth, in more cases rather than fewer in which the standards have application."[47] That is, although the ideal

[45] Schwarzenegger explained in an interview that many aspects of the movie – especially the parts where Schwarzenegger was mean to his competitors – were faked for the sake of making the movie more entertaining. (He explains that the movie was intentionally and explicitly marketed as a "docudrama" rather than a "documentary" for this reason.)

[46] This example is an interesting one to consider in the context of Dover's argument that we do not have an interest in avoiding (hypocritical) criticism.

[47] In comments on a draft of this paper presented at [Removed for Anonymous Review].

application of norms would have Easy apply more demanding standards to *everyone*, there is a less-than-ideal nearby world in which Easy hypocritically applies more demanding standards only to others. This latter world is not ideal, but at least it is better than the actual world.

First note that this objection is compatible with everything I have said. I have attempted an explanation for why a hypocritical set of attitudes is a distinctive kind of bad. I have not made further arguments for any action-guiding upshot of that conclusion. I agree with Arneson that it would be better, all things considered, for Easy to move to the hypocritical possible world from the actual world (and, of course, better still to move to the ideal world). Actual-world Easy's attitudes are not distastefully selfish, but they can be expected to have worse consequences than would Easy's attitudes in the hypocritical-world. Of course, Easy will be criticizable for being a hypocrite in the hypocritical-world, but, again, that is compatible with the objection.

Another way to see the compatibility between the objection and my view is with the following scenario: Suppose Amber judges others (negatively) for being unambitious, but does so with an overly demanding standard of ambition. Amber expects others to be very ambitious (*too* ambitious, objectively speaking). Yet she finds sufficient her own lack of ambition (she is *too* unambitious, objectively speaking). Amber could cease to be a hypocrite by holding herself to the too-high standard or by holding others to the too-low standard.[48] Either option would be as bad as the original scenario, insofar as we are merely evaluating the accuracy of the norms Amber applies (which is all we need to do according to Arneson's objection). But both of those options would be less offensive than the original scenario insofar as Amber's attitudes are no longer influenced by a selfish bias. So there is still an important normative role for hypocrisy to play alongside the injunction to apply norms as accurately as possible.

Section 5 - Conclusion

I have argued for the following analysis of hypocrisy:
Hypocrisy as Selfish Self-Exceptionalism: A person's attitudes or behavior are hypocritical iff she holds others (or out-group entities) to a more demanding standard than that to which she holds herself (or in-group entities) in a case where all else is equal, and does so for selfish reasons.

I have also suggested that the badness of hypocrisy consists primarily in the fact the hypocrite allows a selfish bias to influence her determination of which standards to hold herself to.

[48] Or by picking some other standard to hold everyone too, or even by picking a higher standard for herself than for others

Works Cited

Alicke, Mark, Ellen Gordon, and David Rose. 2012. “Hypocrisy: What Counts?” *Philosophical Psychology*, no. 5: 1–29.

Crisp, Roger, and Christopher J. Cowton. 1994. “Hypocrisy and Moral Seriousness.” *American Philosophical Quarterly* 31 (4): 343–349.

Dover, Daniela. 2019. “The Walk and the Talk.” *Philosophical Review* 128 (4): 387–422. https://doi.org/10.1215/00318108-7697850.

Fritz, Kyle G., and Daniel Miller. 2018. “Hypocrisy and the Standing to Blame.” *Pacific Philosophical Quarterly* 99 (1): 118–139.

Goodman, A. 2007. “Commuting Sentence, Bush Spares Libby from 30 Month Jail-Term.” *Democracy Now*, July 3, 2007.

Hurka, Thomas. 1994. *Principles: Short Essays on Ethics*.

“Hypocrisy (n.).” n.d. In *The Merriam-Webster.Com Dictionary*. Accessed June 10, 2018. https://www.merriam-webster.com/dictionary/hypocrisy.

Kittay, Eva Feder. 1982. “On Hypocrisy.” *Metaphilosophy* 13 (3–4): 277–289.

McKinnon, Christine. 1991. “Hypocrisy, with a Note on Integrity.” *American Philosophical Quarterly* 28 (4): 321–330.

Rai, Tage S., and Keith J. Holyoak. 2014. “Rational Hypocrisy: A Bayesian Analysis Based on Informal Argumentation and Slippery Slopes.” *Cognitive Science* 38 (7): 1456–67. https://doi.org/10.1111/cogs.12120.

Roadevin, Cristina. 2018. “Hypocritical Blame, Fairness, and Standing.” *Metaphilosophy* 49 (1–2): 137–152.

Rossi, Benjamin. 2018. “The Commitment Account of Hypocrisy.” *Ethical Theory and Moral Practice* 21 (3): 553–567. https://doi.org/10.1007/s10677-018-9917-3.

Ryle, Gilbert. 1984. *The Concept of Mind*. University of Chicago Press.

Smith, Angela M. 2007. “On Being Responsible and Holding Responsible.” *Journal of Ethics* 11 (4): 465–484. https://doi.org/10.1007/s10892-005-7989-5.

Soifer, Eldon, and Béla Szabados. 1998. “Hypocrisy and Consequentialism.” *Utilitas* 10 (2): 168.

Szabados, Béla. 1979. “Hypocrisy.” *Canadian Journal of Philosophy* 9 (2): 195–210.

Szabados, Béla, and Eldon Soifer. 2004. *Hypocrisy: Ethical Investigations*. Broadview Press.

“Tractate Bava Batra: Chapter 1.” n.d. Accessed January 8, 2020. https://www.jewishvirtuallibrary.org/tractate-bava-batra-chapter-1.

Wallace, R. Jay. 2010. “Hypocrisy, Moral Address, and the Equal Standing of Persons.” *Philosophy and Public Affairs* 38 (4): 307–341.

The Problem of Self-Sacrifice

Section 1 – Introduction

Could an act of yours count as self-sacrificial even if the result of your action is that which most benefits you? Intuitively, no: self-sacrifice means getting an outcome that sacrifices your own interests. This "no," along with the fact that we often act so as to bring about outcomes we most prefer, is a problem for those who hold that outcomes that most benefit you *are* the outcomes you most prefer.[49]

The philosophers who hold that what most benefits you are the outcomes you most prefer[50] are Wellbeing Subjectivists. So the problem of self-sacrifice is often considered a problem specifically for Wellbeing Subjectivists. To illustrate, consider the following example:

> *Self-Sacrificing Politician:* a politician decides to leave his job in order to take care of his dying parent.

It seems that the Wellbeing Subjectivist will have to say that, insofar as the politician stably prefers the combination of {no job + taking care of parent} to the combination of {job + unable to take care of parent}, the politician has not actually made a sacrifice and so cannot count as acting self-sacrificially. However, intuitively, the politician is acting self-sacrificially (perhaps we would have to add some plausible stipulations about, e.g. how much he enjoys his time under each option to pump your intuitions in this direction). One of the following must be wrong: 1) wellbeing subjectivism or 2) the intuitive extension of "self-sacrifice."

My solution to the problem is external to debates over the proper theory of wellbeing subjectivism and should apply equally well to the problem of self-sacrifice for hedonic and objective list theories of wellbeing. However, the upshot of my solution to the problem is, of course, a defense of wellbeing subjectivism.

We can distinguish between two versions of the problem of self-sacrifice: metaphysical and semantic. The metaphysical problem asks whether it is possible, if wellbeing subjectivism is true, to act against one's own interests (setting aside flawed reasoning, misinformation, etc.) for the sake of someone else.[51] The semantic question asks whether the intuitive extension of "self-sacrifice" can be captured by a Wellbeing Subjectivist. Note that the metaphysical problem is one of mere possibility whereas the

[49] Many philosophers have raised this objection, including: Mark Overvold (Overvold 1980, 117), Richard Brandt (Brandt 1979, Locate Pa#), James Griffin (Griffin 1986, 316), Stephen Darwall (Darwall 2002, 24), and Thomas Carson (Carson 2000, 76). Thanks to Chris Heathwood for this bibliography (Heathwood 2011, 18–19).
[50] Or most value, most desire, take the most subjective interest in, etc.
[51] Thanks to Nicole Hassoun and David Sobel for emphasizing the importance of the metaphysical problem in response to earlier drafts of this paper.

semantic problem is one of extension; in that sense, at least, we can expect the semantic problem to be harder. I resolve the metaphysical question in section 3; as predicted, it is quite a bit easier than the semantic problem. In section 4, I review and criticize several proposed solutions to the problem of self-sacrifice. In section 5, I resolve the semantic problem. In particular, I propose that the everyday definition of "self-sacrifice" has been misunderstood and, once we clear that up, the semantic question is solved.[52] In broad terms, my proposal is that whether an act is self-sacrificial depends on the *motives* for the act, not on the *outcomes* of the act.

Section 2 – Wellbeing Subjectivism and Self-Sacrifice

2.1 – What is Wellbeing Subjectivism?

First, what is wellbeing? A theory of wellbeing helps us identify and explain why a harmful or beneficial state of affairs *counts* as harmful or beneficial (and which cases count as harmful or beneficial). In particular, a person is harmed (or benefitted) when their wellbeing is lowered (or augmented) relative to some relevant comparison class.[53]

I will work with the following schema for subjectivist theories of wellbeing:

> **Unrestricted Wellbeing Subjectivism:** An event or state of affairs, E, is bad (or good) for a person to the extent that that person is disposed to dislike (or like) E.[54]

I call this a schema because "like" / "dislike" are placeholders for positive/negative subjective attitudes a person can hold towards an event.[55] Different wellbeing subjectivists emphasize narrower or wider

[52] Though I should note that Connie Rosati's solution follows the same strategy, albeit with a flawed definition of self-sacrifice, as I argue in Section 4.4. So the originality of my thesis is not in proposing that "self-sacrifice" has been misunderstood, but rather in the specific definition I argue for.

[53] Determining which comparison classes are relevant is also important for deciding what counts as a harm or benefit. But that determination is a separate question than that of wellbeing. The comparison class could be, for example, a temporally prior state, a counterfactually alternative state, etc. (See (Gardner forthcoming) for a comprehensive overview.)

[54] This schema is mostly the same as Eden Lin's "Same-World Subjectivism:"

> "x is basically good (bad) for you at possible world W if and only if and because it satisfies (frustrates) a favorable attitude that you have at W. The extent of x's basic goodness (badness) for you at W is determined by, and proportional to, the strength of the satisfied (frustrated) attitude." (Lin 2019)

[55] Unfortunately, "liking" is awkwardly used when talking about something a person wants to happen for instrumental reasons but does not enjoy of itself (e.g. "I like cleaning my room" is misleading whereas "I want to clean my room" is not). Alternative words I considered have their own problems. "Wanting" is awkward when talking about something that one is already aware of (e.g. "Oh, my room is already clean! I want this to be the case."). Changing the tense ("Oh, good! I wanted this to be the case") doesn't fit well with scenarios where the event is something one hasn't previously considered (e.g. "Oh, I've been knighted? I didn't even know that was a

slices of the pie of attitude-types, e.g. valenced (viz. pro-/con-) attitudes, desires, preferences, values, etc. Sometimes I will talk as if "desire-fulfillment" or "preference-satisfaction" were the going sub-theories – when I do that in this paper, it is merely for linguistic eloquence and not because I am relying on a particular substantive analysis.

2.2 – The Problem of Self-Sacrifice (In More Detail)

In formulating the problem of self-sacrifice, Heathwood proposes the following relevant principle:

> "A Principle about Welfare and Self-Sacrifice: An act is an act of self-sacrifice only if the act fails to be in the agent's best interest." (Heathwood, 2011, page 21)

Wellbeing subjectivism makes a substantive claim about the concept "an agent's best interest:" best interests are constituted by the outcomes that the agent is disposed to like the most. This gives us the following corollary:

> "A Principle about Wellbeing Subjectivism and Self-Sacrifice: An act is an act of self-sacrifice only if the act fails to produce the outcome the agent is disposed to like the most."

We can now illustrate the problem of self-sacrifice more precisely. The first example is due to Heathwood (Heathwood 2011, 32):

> "*Alice's Friday Night:* Alice is deliberating over how to spend her Friday night. She can go to the disco with her friends, or she can volunteer at the soup kitchen. Alice considers the options and, despite how badly she wants to go dancing with her friends, she decides, voluntarily and with full and vivid knowledge, to spend her Friday night helping the needy at the soup kitchen. She feels it would be the right thing to do, and so she does it."

Intuitively, Alice acts self-sacrificially. So Alice's actions fall within the ordinary extension of "self-sacrifice." Now let us ask the relevant question: under the assumption of the truth of wellbeing subjectivism, did Alice's action (viz. volunteering at the soup kitchen) benefit her more than the relevant alternatives (viz. going to the disco)? Yes: the degree to which Alice is disposed to like volunteering at the soup kitchen is higher than the degree to which Alice is disposed to like disco (on this particular Friday night at least). That means that the Principle of Wellbeing Subjectivism and Self-Sacrifice has been violated. (One might object here that Alice chose to volunteer out of a sense of duty rather than

---

possibility. Anyway, I wanted this to happen!" is very confusing, whereas "Oh, I like that this happened!" is not). "Prefers" is problematic in that it is always comparative – ideally our theory of wellbeing gives us absolute values.

because of the sort of positive attitudes that contribute to wellbeing. I consider this suggestion in section 4.4. I also consider Heathwood's own analysis of *Alice's Friday Night* in section 4.6.)

The second example is inspired by the character of Thomas Pembridge from the television series *Mozart in the Jungle*.

> *Thomas the Conductor:* Thomas has been the conductor of the New York Symphony for many years and his skills are beginning to decline in his old age. He deliberates over whether to retire or continue conducting. He decides, *for the sake of the music*, to retire.[56] (He reasons that the music produced by the symphony will be better if the up-and-coming conductor, Rodrigo, takes over.) Thomas is disposed to have a stronger positive attitude towards the state of affairs in which the symphony's music is better than the attitudes he would have towards the state of affairs in which Thomas continues to conduct (even though Thomas would enjoy conducting for many more years).

Thomas has made a self-sacrificial choice – he values the music even at a cost to his own enjoyment of being the conductor. On the other hand, unrestricted wellbeing subjectivism tells us that Thomas's retiring also benefitted him more than the alternative since he is disposed to prefer the state of affairs in which the music benefits most *over* the state of affairs in which he continues conducting. Again, the Principle of Wellbeing Subjectivism and Self-Sacrifice has been violated.

Section 3 – The Solution to the Metaphysical Problem

The metaphysical problem of self-sacrifice is solved if we can show that it is *possible* to act against one's own interests for the sake of someone else.[57] Note that, in contrast to the semantic problem, it does not depend on the meaning of "self-sacrifice."

A hedonist theory of wellbeing has a very easy time solving this problem: find a scenario in which someone makes a decision that they know will make them unhappy. Wellbeing subjectivists must find a scenario where someone makes a decision that leads to an outcome they do not desire. That sounds a little strange but if we focus on a case where the agent knows they will have a change of desire – and

[56] Importantly, the music is an end in itself, not an instrumental end for, e.g., Thomas's listening pleasure.
[57] I presume that we are meant to set aside cases where the agent acts irrationally, out of weak-will or akrasia, or due to misinformation or ignorance.

given that it is common for people to discount their future desires (or even happiness) when compared with their present desires – the scenario should not seem so strange.

Analyzing the wellbeing impact of a scenario where someone's desires change is somewhat complicated, since there are four different views of when a desire-satisfaction is supposed to benefit an agent. With respect to different possible timings of benefit, Ben Bradley writes:

> "Suppose S desires that P. Suppose the desire happens at time t1, and P obtains at time t2. When, if ever, is S benefited by this? There are four answers that have been defended:
>
> at t2 only (the "time of object" view);
> at t1 only (the "time of desire" view);
> at only whichever of t1 or t2 is later (the "later time" view);
> at t1 *and* t2, if t1=t2; otherwise at no time (the "time of both" view)." (Bradley 2016)

Bradley defends the "time of both" view (see also (Heathwood 2005), (Heathwood 2011); c.f. (Lin 2017)). Based on Bradley and Heathwood's arguments, the time of both view seems to me the only plausible one. In any case, let us see how each view fairs with respect to the metaphysical problem of self-sacrifice. I will use the same case for each of them:

> *Unstable Preferences Politician:* Jane, a politician, currently desires to spend less time working and more time with her children. (She also knows that her children want her to do this.) However, she knows that, in a month, she will desire to spend more time working and less time with her children. If she abandons her political position now, she will be unable to reverse course later. She decided to abandon her political position now.

Time of Object View: The time of object view locates benefits when the relevant event occurs. So let us fast forward a month and see what happens. Jane is spending more time at home with her children but now wishes she was back in the office. Regardless, she is *currently* benefiting from spending more time at home because her desires a month ago have been fulfilled. The "object" of her past desire is *currently* being fulfilled – it is the time at which this "object" occurs that determines when the benefit occurs. Of course, Jane also has a current desire to be back in the office, and the object of that current desire is right now. So that desire-frustration is a harm. I will just stipulate that the current desire is stronger than the past desire. Therefore although Jane is currently benefiting from the satisfaction of her past desire to leave the office, she is currently being harmed *to a greater extent* by the frustration of her current

desire to return to the office. Therefore, she has managed to sacrifice her own wellbeing for the sake of her children, and therefore the metaphysical problem of self-sacrifice is not a problem for the Time of Object View.

Time of Desire View: The time of desire view locates benefits when the desire occurs. So if Jane now desires to leave her job (and she does eventually do so) she will benefit now. Even if the object of her desire does not occur for another month, she benefits *now,* at the time of her desire. Of course, a month later, Jane will have the desire to return to the office. And so her desire-frustration in a month will constitute a harm. As before, I will stipulate that her desire in a month is stronger than her desire now. So, again, Jane's decision will, overall, reduce her wellbeing.

Later Time View: The later time view coincides with the object of desire view in those cases where the object of one's desire occurs after the desire itself. That is how things are in *Unstable Preferences Politician*, so my analysis of the object of desire view applies here.

Time of Both View: According to the time of both view, only desires concurrent with their objects are relevant to wellbeing. Desires about what occurs in the future or about what occurred in the past are not directly relevant. After Jane leaves her job, her desire to be at home is satisfied every moment for a month. Then her desire reverses, and her desire is frustrated thenceforth. I will stipulate that her desire in a month is stronger (and lasts for a longer amount of time) than her initial desire. So Jane has managed to sacrifice her wellbeing.

Therefore *Unstable Preferences Politician* (with the right stipulations) solves the metaphysical problem of self-sacrifice regardless of when desire-satisfactions benefit. I turn next to the semantic problem.

Section 4 – Problems for Previously Proposed Solutions to the Semantic Problem

4.1 – Benevolent Preferences Excluded

One straightforward attempt to solve the problem of self-sacrifice is by excluding the satisfaction of benevolent or moral preferences from the calculation of an agent's wellbeing. Either of these exclusion rules would give us the right answer in *Alice's Friday Night*: Alice is no longer benefited when she helps the soup kitchen's patrons, no matter how favorable an attitude she takes towards that event, because her preference to help them is benevolent and moral. Therefore, Alice would have benefitted more from going to the disco and therefore Alice does act against her best interests by not going to the disco.

Excluding moral or benevolent preferences results in an implausible theory of wellbeing, however. I understand "moral preferences" to be a narrower domain of preferences than "benevolent" ones. Both kinds of preferences involve acting for the sake of *something* other than oneself. That *something* could be anything with a benevolent preference, but, depending on one's theory of morality, might require a narrower range of targets to qualify as a moral preference. (E.g. a welfarist will only count preferences that aim to benefit wellbeing-subjects as moral preferences but could still count a preference to benefit, say, Pluto, as benevolent.)

I will start with a critique of excluding moral preferences, or "Morals Excluded" for short, before turning to "Benevolence Excluded." The "morals" in Morals Excluded might refer to subjective moral aims or objectively moral ones. A subjective moral aim is one in which the agent truly believes that their aim is moral. An objective moral aim is an aim that *is* moral, even if the agent believes otherwise. Neither version of Morals Excluded is a plausible theory of wellbeing however.

If Morals Excluded refers to the exclusion of *subjectively* moral aims, on the one hand, then it cannot account for the self-sacrificial act of Thomas, since, for Thomas, acting for the sake of music is an aesthetic aim, not a moral one. Furthermore, it will have problems when an agent makes what they *think* is a non-moral choice. In the well-known case of Huckleberry Finn, the agent acts for the sake of his friend, Jim's, wellbeing, despite believing that he is acting wrongly. Subjective Morals Excluded would not exclude the satisfaction of Huck's desire to help Jim from benefiting Huck, and thus would not solve the semantic problem of self-sacrifice for this kind of case.[58]

If Morals Excluded refers to the exclusion of *objectively* moral preferences, on the other hand, then self-sacrifice for political causes could only occur if the political cause were objectively just. No one on the wrong side of history would qualify as a self-sacrificer! Nor does anyone who acts for the sake of an objectively irrelevant cause get to be a self-sacrificer, either (here I imagine someone spending many hours of their life canvassing for a political candidate who differs only insignificantly from the opposing candidate). Thus the dilemma: whether we exclude only subjective moral preferences or only objective ones from the wellbeing formula, we arrive at an implausible theory of wellbeing.[59]

The exclusion of moral preferences has other problems as well, unrelated to the problem of self-sacrifice. Consider the following example:

[58] Thanks for Julia Markovits for pointing out the applicability of this case.

[59] Sobel also criticizes the inability of the moral-preferences-excluded-formula to solve the problem of self-sacrifice for the following reason: actions taken for the sake of a group one identifies with – e.g. one's nation, religion, team, etc. – can intuitively be self-sacrificial without being moral. (Sobel 2016)

> *Fictional Concern:* I hear a story about an innocent person who has been framed for murder. I am unsure whether I am being told a fictional story or a non-fictional one. I decide that in either case I want the protagonist to be found not guilty.

Morals Excluded claims, implausibly, that I am benefitted or harmed by the protagonist's fate *only if* the protagonist is fictional but not if the protagonist is real.

One might object that if I am indifferent to whether or not the protagonist is real then my preference must not be a moral one. I could change the example slightly to respond to this objection: imagine instead that there are two different people, each of whom hear the same story and develop the same preference, but the first person assumes the story is fictional whereas the second person assumes the story is non-fictional. I still have the intuition that, if one of them can be non-instrumentally harmed by the outcome of the story, then both of them can. Or if, instead, one of them cannot be non-instrumentally harmed by the outcome of the story, then neither of them can.

One could also object that concern for the fate of the fictional protagonist *is* a moral sentiment, contra my assumptions, and so *is* excluded from a subject's wellbeing calculus by Morals Excluded. But this objection entails that my wellbeing can be affected by the success of a fictional character for whom I care merely because I find the character attractive (that is, I relate to and therefore sympathize with the character for non-moral reasons), but my wellbeing cannot be affected by the success of a character for whom I empathize out of a recognition of her moral status. Or suppose I wanted the fictional protagonist to be found innocent not out of empathy but because of the aesthetic demands of the story's plot. Intuitively, if I can be benefited by the fate of a fictional character because of a preference I have, then I can be thusly benefited whether that preference is moral or aesthetic.

Another problem arises in cases of moral, self-interested desires. If an agent is in a situation where the moral thing to do is also the action that benefits themselves the most then, according to Morals Excluded, they have not benefited. That is clearly absurd.[60] For example:

> *The Wise Carpenter:* A wise carpenter reasons that hitting the nail, rather than his finger, with a hammer is both the most moral thing to do and the most self-interested thing to do. He chooses to hit the nail.

[60] Thanks to Julia Markovits for suggesting this problem.

So much for Morals Excluded, then. Excluding benevolent preferences seems more promising, so long as we take a wide view of what counts as benevolent. "Benevolence," as I am using it, means an intention to benefit other entities, persons or otherwise. But, although music is not a person, there is a sense, I believe, in which Thomas the Conductor views music as the intended beneficiary of his actions. A beneficiary is something that is benefited and so I find it natural to use the word benevolence here. (Whether or not music is *objectively* the kind of entity that is capable of being a beneficiary, it is at least capable of being a beneficiary *subjectively*). So "benevolence," in my usage, means an intention or desire to make some *thing* (other than oneself) better for the sake of that thing (i.e. one must implicitly view the beneficiary as an end-in-itself). Excluding benevolent preferences from the wellbeing calculus does draw the line in the right place so far as solving the problem of self-sacrifice is concerned. But it is an ad hoc solution and, as with most ad hoc solutions to narrow problems, we can expect it to generate its own counterexamples.

Consider the following scenario:

> *George, Immanuel, and the Cherry Tree*: Suppose Immanuel and his parents have just returned home and discover that their cherry tree has been cut down. His parents ask their other son, George, if he knows what happened. Immanuel wants George to tell the truth.

Whether or not Immanuel is harmed by George lying, according to Benevolence Excluded, depends on the aims that ground Immauel's desires. If Immanuel wants George to tell the truth because Immanuel does not like lying in general, then Immanuel's desires are not aimed at benefiting anything – they are not benevolent. So Benevolence Excluded allows George's impending action to non-instrumentally benefit or harm Immanuel. If, instead, Immanuel wants George not to lie because Immanuel is concerned for George's wellbeing (Immanuel knows that if George lies it will only make George's eventual punishment worse), then Benevolence Excluded does not allow for George's pending action to benefit or harm Immanuel. Intuitively, Immanuel should be benefitted by George's honesty regardless of why he has that desire.

4.2 – Non-remote Wellbeing Subjectivism

Another popular modification to Wellbeing Subjectivism carves out a much wider exception than Moral or Benevolent preferences: exclude preferences concerning remote states of affairs. A state of affairs or event is remote to a person when it is not about or does not involve that person (other than for the fact that the person has a subjective attitude toward it). Writers on the topic of unrestricted wellbeing subjectivism have argued in favor of excluding remote desires on grounds independent to the problem of self-sacrifice. So the remote desires exclusion rule has at least one thing going for it: it is not ad hoc with respect to the problem of self-sacrifice.[61] Indeed, the exclusion rule *sounds* right to me: states of

[61] The most popular example in the literature is due to Derek Parfit: "Suppose I meet a stranger who has what is believed to be a fatal disease. My sympathy is aroused, and I strongly want this stranger to be

affairs that are capable of benefitting or harming me should be *about* me or *involve* me in some relevant sense. But then I think "wait, do I not involve myself in a state of affairs by taking a valenced attitude towards it?"

In any case, accepting the proposed modification here yields:

> **Wellbeing Subjectivism, Non-remote:** A *non-remote* event or state of affairs, E, makes a person worse off (or better off) to the extent that that person is disposed to dislike (or like) E.

So, under Non-Remote, Alice does not directly benefit when she increases the wellbeing of the soup kitchen's patrons since the wellbeing of the soup kitchen's patrons are remote with respect to Alice.[62]

What exactly is it for a state of affairs to be non-remote, viz. to be about or involve a person? Overvold has proposed that a preference counts as remote if the preferred state of affairs could obtain despite the non-existence of the subject.[63] For example, if I prefer that North Korea disarms its nuclear weapons, this preference would count as remote since it is possible for North Korea to disarm even if I cease to exist. In contrast, a preference for the flavor of chocolate ice cream is non-remote since I can only fulfill a preference for this flavor while I exist.

One problem with Non-Remote Wellbeing, raised by David Sobel, is that it does not adequately account for self-sacrifice by agent-centered/deontologically motivated actions. (Sobel 2016) Consider, for example:

> *George and the Cherry Tree:* George's parents discover that their cherry tree has been felled and ask George what happened. George tells the truth and admits his guilt. He acts for the sake of being honest even though he believes this will incur a significant cost.

So long as George is acting for the sake of *his own* honesty (rather than out of concern for honesty in general[64]), then the state of affairs he prefers is non-remote (both intuitively and by Overvold's criterion). Since his own honesty is non-remote, he will be benefitted by the fact that he did not lie. With the right stipulations, it will turn out that not lying was both self-sacrificial and, according to even Non-

---

cured. We never meet again. Later, unknown to me, this stranger is cured. On the Unrestricted Desire-Fulfilment Theory, this event is good for me, and makes my life go better. This is not plausible." (Parfit 1984, 494)

[62] Unless, as Julia Markovits has pointed out, in conversation, Alice's desires that it be *because of her* that the patron's benefit.

[63] (Overvold 1980). Technically the object of the preference is what is remote, but I call the preference itself remote for the sake of concision.

[64] The distinction can be illustrated by asking whether an agent would prefer a world in which he lied or one in which he told the truth but five other people lied.

Remote Wellbeing, in his best interests. Applying Sobel's point to this example: intuitively, George is acting self-sacrificially; yet none of the satisfied preferences are remote. So the semantic problem of self-sacrifice has not been fully addressed by excluding remote preferences.

On the other hand, my intuitions about cases involving agent-centered motivations are unclear. George's motivations are, to use Bernard William's phrasing, "morally self-indulgent" (B. Williams 1981, chap. 3), such that I am not sure that I want to count his action as intuitively self-sacrificial. He is not acting for the sake of honesty, he is acting for the sake of his *own* honesty. In acting to preserve his own virtue, George strikes me as someone who acts merely for himself (since I assume this kind of person, following Aristotle, thinks that being virtuous directly increases one's wellbeing). That is not the kind of benevolent or other-regarding motivation that I take self-sacrifice to require. Granted, these are the kind of intuitions that tend to be distinct to consequentialists, so they may not be widely shared. I return to the question of moral self-indulgence and self-sacrifice in relation to my positive proposal, in section 5, below.

In any case, Non-Remote Wellbeing faces other problems apart from whether it adequately solves the problem of self-sacrifice. A rule that excludes certain desires is shown to be artificial if there are cases where the satisfaction of Preference A and Preference B seem, intuitively, to be on a par for affecting Subject P's wellbeing, yet the exclusion-rule draws a hard line between them. In other words, the counterexample schema contains two preferences, A and B, for which the following are both intuitively true: 1) Preference B is more plausibly (directly) wellbeing-affecting than Preference A yet 2) Preference B is more remote then Preference A.

The following counterexamples rely on the intuition that a preference for success or failure in a project one has contributed to is non-remote. Suppose I work for a law firm that allows each of its lawyers to work on only one of the law firm's pro bono projects at a time. Lawyers without a project choose which project to work on in order of seniority. I and one other lawyer are currently unassigned to any pro bono projects and there are currently two projects available. Since I have seniority, I will get to choose which project to work on, while my colleague will be stuck with the other one. Suppose I really care about the outcome of Project A (for idiosyncratic and non-instrumental reasons) but, since my colleague is a specialist in the field of law relevant to that project, I leave it to her. Instead I take on Project B, whose outcomes I care little about. Non-remote Wellbeing Subjectivism has the unintuitive result that the success of my own project, being non-remote, benefits me, but the success of my colleague's project does not, even though I care much more about the latter.

Other examples can be found to fit the schema, some of which I present here in tabular format (the first row after the heading expresses the schema in a generalized format and the remaining rows instantiate the generalization):

Preference A (relatively non-remote)	Preference B (relatively remote)	Preference B is more plausibly wellbeing-affecting than Preference A	Preference A is less remote than Preference B
The success of a project I am involved in (as an agent)	The success of a project I am *not* involved in, though one whose success I care more about	✓	✓
A weak preference to succeed in an inconsequential task my employer has assigned me	A strong preference for my co-worker's immensely important project to succeed	✓	✓
A weak preference to win a pick-up game of basketball	A strong preference for a team I am not on to win their match	✓	✓
A police officer's weak desire to be the one to save my neighbor's cat from a tree	My strong preference that my neighbor's cat be saved	✓	✓
A weak preference to play the violin well	A strong preference for a person - whom you used to tutor in philosophy - to succeed in his ambition to play violin professionally.	✓	✓

One might object that the remote-preference-satisfactions seem intuitively beneficial because of the pleasure that is correlated with satisfying one's desires. If that were true, it would undercut the strength of these counterexamples. To clarify: although a Non-Remote theory of wellbeing would *not* say that the remote-preference-satisfactions are *directly* beneficial, it could say that they are non-directly beneficial insofar as they cause pleasure. To avoid this objection, we could modify all of the scenarios above such that the patient never finds out if their preferences are satisfied. Thus the patient does not experience any pleasure or displeasure resulting from their preference-satisfaction, which may avoid the problem of distorted intuitions.

4.3 – Behavioral-Desires Excluded

Heathwood has recently proposed excluding desires in the behavioral inclination sense (while including desires in the genuine appeal sense) from the wellbeing calculus, echoing Parfit's response to psychological egoism. Unfortunately, as I will argue, the appeal to this distinction does not succeed at

addressing the problem of self-sacrifice, since many self-sacrifices are made with desires in the genuine appeal sense.

To introduce this distinction, I will take a detour through Parfit's comments on psychological egoism. By "Psychological Egoism," Parfit means the view according to which everyone always acts selfishly. Parfit summarizes a relevant argument for this view as follows: "Whenever people act voluntarily, they are doing what they want to do. Doing what we want is selfish. So everyone always acts selfishly" (Parfit 2011, 43). Parfit claims the argument equivocates in its use of "want" and is therefore invalid.

The two senses of wanting that the psychological egoist equivocates are: 1) being merely inclined to cause the object of desire to come about and 2) finding the object of desire genuinely appealing.[65] Tamar Schapiro explains the distinction quite clearly:

> There is one sense of "desire" or "want," such that whenever you act (where the idea of action implies that it was in some sense free, intentional, voluntary, etc.), we can say you had a "desire" to do what you did. In this sense of "desire," it is logically impossible to do something without "having a desire" to do it. To attribute a "desire" in this sense is just to attribute motivation to the agent, as the conceptual correlate of action. But there is another sense of "desire" or "want," that allows for the possibility of doing something without having a desire to do it. When you take out the garbage even though you do not feel like taking out the garbage, you do something even though you have no desire, in the second sense, to do it. You lack a certain kind of motivation. But we can still attribute to you a desire to take out the garbage, in the first sense." (Schapiro 2014, 136)[66]

Returning now to the argument for psychological egoism, the thought is that we can grant that "whenever people act voluntarily they are doing what they want to do" but only if "want" includes *both* the behavioral inclination and genuine appeal sense of the word. The next premise claims that "Doing

[65] Heathwood identifies attentiveness to this distinction in a wide array of work on the philosophy of desire/wellbeing: (Daveney 1961), (Nagel 1970), (Foot 1972), (Lewis 1988), (Sumner 1996), (Campbell 2013), (Schapiro 2014), and (Hume 1739). And, of course, Heathwood himself elaborates on the distinction (Heathwood 2017, 11–12). See also: (Schueler 1995).
[66] David Sobel summarizes the distinction as: "Oh boy, I get to" vs. "Oh, I got to" attitudes. (In conversation.)

what we want is [always] selfish;" but this is only plausible if "want" is being used in the genuine appeal sense of the word. Since different senses of the word "want" are at play, the argument is invalid.

So much for psychological egoism; does a similar train of thought deliver a solution to the problem of self-sacrifice? The suggestion, more precisely, is to define self-sacrifice as follows:

> Parfit-Inspired Definition of Self-Sacrifice: an act is self-sacrificial if a person acts on an action-desire and out of benevolent motivations.

Note that the solution offered by the Parfit-Inspired Definition of Self-Sacrifice is not internal to the philosophy of wellbeing. It is a self-contained conceptual analysis of "self-sacrifice." Of course, wellbeing subjectivism could be restricted to exclude action-desires – which would be internal to the philosophy of wellbeing. In fact, Heathwood makes just this proposal:

> "My aim in what follows is to show that if, for the purposes of the desire theory of welfare, we understand 'desire' in the ordinary, attitudinal, true, affective, inclinational, warm, appetitive, violent sense rather than in the merely behavioral, intentional, volitional, non-affective, cold, calm, wide, philosophers' sense, we can provide plausible solutions to [several problems for the desire theory of wellbeing]." (Heathwood 2017, 14)

In any case, Heathwood's proposal works as a solution to the problem of self-sacrifice *by making* the Parfit-Inspired definition derivatively true, though one could also just accept the Parfit-Inspired definition outright without any particular theory of wellbeing in mind.

However, I do not find that the Parfit/Heathwood proposals square with my intuitions about particular cases of self-sacrifice and so am skeptical that the semantic problem has been solved. Consider the following case (based on a true story):

> *Harriet the Graduate Student:* Harriet, a graduate student in philosophy, has barely any income. But she is so passionate about the plight of destitute persons in impoverished countries that she gives her entire life savings (several thousand dollars) to a famine-relief charity.

I think it is quite plausible (and somewhat commonplace) for people to find helping others genuinely appealing. The Parfit/Heathwood solution says 'well, since she was passionate about her act of charity, it does not count as self-sacrifice. She would have had to do it begrudgingly or coolly to qualify.' That seems wrong to me.

Similar problems face the Parfit/Heathwood solution when applied to *Alice's Friday Night* or *Thomas the Conductor*. We can stipulate that Alice does not work at the soup kitchen dispassionately – that she

finds it genuinely appealing. Likewise, we can stipulate that Thomas is not merely behaviorally inclined to retire for the sake of the music – he wants the music to be as good as possible in the genuine appeal sense of want. I have the intuition that both Alice and Thomas are still acting self-sacrificially.

Another problem arises in cases where the agent lacks foreknowledge. Suppose Alice is not passionate about the soup kitchen when she decides to spend her evening there, but while working there she reunites with her long-lost twin sister and her life goes much better as a result. Intuitively, her decision to work at the soup kitchen was self-sacrificial. But Heathwood's Principle about Welfare and Self-Sacrifice ("An act is an act of self-sacrifice only if the act fails to be in the agent's best interest") has obviously been violated. (I return to this problem of foreknowledge in motivating my own solution, in Section 5.)

So although Heathwood's Behavioral-Desires Excluded theory of wellbeing may be independently correct (for other reasons[67]) it does not solve the problem of self-sacrifice. (Though it does *help* with the problem, insofar as excluding behavioral-desires fixes *some* of the extensional mismatch between intuitive self-sacrifice and self-sacrifice in light of wellbeing subjectivism.)

4.4 – Rosati's Solution

Like the Parfit-inspired solution (above) and my solution (below), Connie Rosati also argues that an act can be self-sacrificial *and* in one's own best interest (Rosati 2009). Rosati's proposes that an can be self-sacrificial if the cost the agent incurs is one that involves either a physical sacrifice or a pursuit, activity, or relationship with which the agent is deeply engaged.

Rosati asks us to consider the three paradigm types of self-sacrifice: sacrifices of life, limb, or love. What unites these paradigm cases? Rosati argues that all three involve risk of harm to one's "self." Rosati uses "self" in a broad, technical sense which includes not only the physical body, but also projects, activities, and relationships with which one is deeply engaged.[68] To be deeply engaged with a project, activity, or relationship is for that project (etc.) to be something the agent "does not merely enjoy or feel glad for but loves,"(Rosati 2009, 318) to reward her (or for the other member(s) of the relationship to love her back), "to support [her] sense of her own worth and provide a sense of direction and identity…[to be] experienced as internally motivating rather than as forced on her from outside."(Rosati 2009, 318) Rosati also says "In loving something, we come to give it an organizing position in our lives; our loves become the things around which we arrange our time and our other activities and engagements…They also help to determine our identity, our views about who we are and what our lives are about…"(Rosati

[67] And, in fact, I think it *is* independently correct - see my "The Problem of Non-Sentient Robots" (unpublished).
[68] This technical usage of "self" makes me suspect that Rosati's insight about paradigm cases of self-sacrifice is merely a concealed tautology.

2009, 318) Someone who risks their life is thereby risking "all of the projects, activities and relationships that are also…parts of her good intimately connected to her self."(Rosati 2009, 317) Likewise, Rosati argues, for risks to one's limbs: "To risk one's physical integrity is thus, as when one risks one's life, to risk those projects, activities and relationships that are also connected to the self."(Rosati 2009, 317) And so on.

These considerations lead Rosati to propose the following:

> "In order for an act to be an act of self-sacrifice, it need not involve a net loss of welfare considering one's life as a whole. And it need not involve trading one's good for some value which is not a part of one's good; it can, instead, involve sacrificing one part of one's good in favour of another part, provided that one does so out of a regard for the value or good of another. To be an act of self-sacrifice, however, it cannot involve the sacrifice of just any benefit. Rather, *it must involve the sacrifice of some part of one's good that is at the same time a sacrifice of self.*" (Rosati 2009, 319)

I have emphasized the last sentence in this quote since that is where Rosati and I diverge.

Rosati's solution works well for the *Thomas the Conductor* case: actively conducting *is* a central part of Thomas's life in exactly the sense Rosati intends. And, although Thomas's retirement is in his own best interests (since it is the way to get what he most wants), his action will still count as self-sacrificial. The solution does not work well for *Alice's Friday Night*: Alice is giving up an evening of fun with her friends. We can stipulate that her relationship with her friends will not be lessened in anyway by forgoing a single party with them; so no sacrifice to her loving relationships occurs. Rosati's solution also does not work well for *Harriet the Graduate Student*, for whom a loss of her life's savings is not a loss of anything having to do with her sense of self, nor any project that she is deeply engaged with (if necessary, we can stipulate that Harriet's savings were disposable income). Where Rosati's reasoning goes astray is her identification of the paradigm types of self-sacrifice. Sacrificing pleasure or money are *also* paradigm ways to act self-sacrificially, yet neither of these kinds of self-sacrifice will typically rise to the level of a sacrifice of "self" in Rosati's sense of the word.[69]

[69] Although perhaps sacrifices of money lead, instrumentally, to lessened ability to complete one's own projects. (Thanks to Abigail Bruxvoort for this suggestion on Rosati's behalf.)

4.5 – Not Self-Sacrifice?

Heathwood argues that, once we get clear about all the details of many alleged self-sacrifice counter-examples, we will realize our initial intuitions were mistaken; we will realize that, in fact, no self-sacrifice occurs. Consider again Heathwood's example, which I reprint here for convenience:

> *"Alice's Friday Night*: Alice is deliberating over how to spend her Friday night. She can go to the disco with her friends, or she can volunteer at the soup kitchen. Alice considers the options and, despite how badly she wants to go dancing with her friends, she decides, voluntarily and with full and vivid knowledge, to spend her Friday night helping the needy at the soup kitchen. She feels it would be the right thing to do, and so she does it." (Heathwood 2011, 32)

Heathwood argues: "In order for this alleged counterexample to work, the following must hold:

- In spending her evening at the soup kitchen, Alice is getting what she most wants on this evening;
- Alice will not lose out in the future on things she will be wanting in the future by going to the soup kitchen tonight…;
- Had Alice gone to the disco instead, she would have, during her whole time there, been fairly strongly wanting to be at the soup kitchen.
- Had Alice gone to the disco instead, she would not have formed all manner of new desires for what befell her at the disco, or had a very strong desire to be there at the disco."(Heathwood 2011, 34)

The third and fourth bullet points ensure that Alice would not have benefited *more* from going to the disco than she does from going to the soup kitchen. For, if it were true that, although Alice chose to go to the soup kitchen she *would* have benefited more by going to the disco, then it turns out that Alice is making a global sacrifice to her wellbeing that Wellbeing Subjectivism can already accommodate (see action 3, above).

Once we have all these facts of the case vividly presented to us, Heathwood argues, we will see that it really *is* in Alice's best interest to go to the soup kitchen, and we can have this intuition even if we were not already committed to wellbeing subjectivism. And, since it is in her best interest, it must not be an instance of self-sacrifice.

I agree with Heathwood that it is in Alice's best interests to work at the soup kitchen. When someone brings about the outcome they *will* most want to occur, then they *do* benefit most. But my intuitive sense is that Alice is still acting self-sacrificially…somehow. Thus it seems that my intuitions about cases go against Heathwood's Principle about Welfare and Self-Sacrifice. According to that principle, "An act is

an act of self-sacrifice only if the act fails to be in the agent's best interest."(Heathwood 2011, 21) In section 5 I argue against this principle and conclude that Alice has acted in her best interest *and* acted self-sacrificially.

Section 5 – The Solution to the Semantic Problem

I argue that the problem of self-sacrifice arises from a misunderstanding of what that concept actually requires. Contrary to what many philosophers have claimed, acts *can* be self-sacrificial despite being most beneficial to the actor, *so long as the actor has the right motivations*.

To start with, here are two examples to support the claim that the ordinary concept of self-sacrifice accommodates acts that most benefit the actor (even in the absence of a subjectivist theory of wellbeing):

1) *Fidel the Revolutionary*: Fidel joins a revolutionary militia group. He suffers many hardships during the war. When the war ends, and because of his many acts of heroism, he becomes the leader of his country and his life goes very well for him. His life goes so well that it counterbalances the harms he faced during the war. Intuitively, Fidel's decision to join the revolution was both self-sacrificial and, at the end of the day, in his best interest.[70]

2) *Mamoudou Gassama*: Mamoudou was walking down the street in Paris when he saw a young child dangling from a balcony railing, four stories up. He quickly climbed the outside of the balcony – risking personal injury – and rescued the child. The president of France took notice of the widely circulated video of Mamoudou rescuing the child and granted French citizenship to Mamaoudou. (Mamoudou had recently immigrated from Mali.)

[70] Assume that, counterfactually, he would have had a boring life of mediocrity if he had not joined the revolution.

> Thus Mamoudou's act was self-sacrificial and, in the end, in his best interest.[71]

What features of these cases give rise to the surprising result that a self-sacrificial act can most benefit the actor? The answer is this: self-sacrifice, unlike "most benefits," is partly determined by a person's motivations at the time they make their decision. We judge that Fidel and Mamoudou act self-sacrificially because they are selflessly *motivated* (for the sake of a political cause in Fidel's case, for the wellbeing of a stranger in Mamoudou's). Whether a state of affairs most benefits a person, in contrast, is determined by how well things actually go for a person, regardless of that person's motivations.

Each of these cases has an additional feature, however, that sets them apart from the other examples discussed earlier: the actors (Fidel and Mamoudou) are unaware that their actions will most-benefit them. In contrast, Alice the Soup-Kitchen Volunteer, Thomas the Conductor, and Harriet the Effective Altruist are not relevantly ignorant: they know all the relevant consequences of their respective choices. If the examples of Fidel and Mamoudou are to lend persuasive force to my argument, I will need to argue that an actor who *knows* that an act will be most self-beneficial can still act self-sacrificially. Call this the "Problem of Foreknowledge."

In my analysis of Fidel and Mamoudou, I claimed that each actor's motivations did the work in qualifying their respective actions as self-sacrificial. That each actor did not know that their chosen action would most self-benefit provides us, the reader, with clear intuitions that these actors were selflessly motivated. So my argument is as follows: paradigm examples of self-sacrifice + most-self-benefit will involve actors who strike us as being unquestionably selflessly motivated. Since an actor who is ignorant of the fact that his act will be of most self-benefit is unquestionably selflessly motivated, ignorant actors will best serve as paradigm intuition pumps. But the ignorance of these actors is not directly relevant; their ignorance is relevant only for pumping our intuitions regarding their stipulated motives. Instead, the actors' motivations are what qualify their acts as self-sacrificial.

In support of the claim that motivations, but not foreknowledge, is what is pivotal for self-sacrifice, I will first argue that motivations are detachable from foreknowledge in general. Then I will introduce some modified versions of *Fidel the Revolutionary* in which we see what kind of difference it makes if we change the actors motivations (but keep fixed their foreknowledge) versus changing their foreknowledge (but keeping fixed their motivations).[72]

---

[71] Everything in this case actually happened (in May, 2018): https://www.cnn.com/2018/05/28/asia/paris-baby-spiderman-rescue-intl/index.html

[72] One might wonder why I even bother with ignorance-cases in the first place. The answer is that, as I explained, such cases are best at drawing out the intuition that an act can be self-sacrificial and most benefit the actor.

That motivations can be unaffected by (extra) relevant beliefs occurs in non-moral situations as well. Imagine a professional tennis player, for example, that plays tennis simply because she enjoys the game. She is well aware of the fame, glory, and money that she will accrue if she wins a major tournament, but it is still possible that *this* tennis player is not motivated by those goals. Imagine I presented you with two cases, in each of which a tennis player played her hardest in every tournament she could. One of the tennis players, however, always finishes near the bottom of the tournament pool – and she knows this, suffering no delusions. The other player usually wins first place. It is possible that each player has the same motivations: the mere joy of playing tennis. But even if I stipulated that this were the case, I suspect that your intuitions would not be fully "sold" on my stipulation; you would suspect that the winning player *must* be motivated by something more. Indeed this is a common problem that arises when we stipulate the occurrence of unusual and counterintuitive features in our thought experiments. Intuitions form based on preconceived associations – they often do not give our stipulations their full due.

Returning now to the case of *Fidel the Revolutionary*, we can perform a few thought-experiments with "control groups" to see whether it is foreknowledge or motivation that really makes the difference. Here are two alternatives to *Fidel the Revolutionary* (in each description, I italicize the changes that have been made from the original):

> *Selfish, Ignorant Fidel*: Fidel joins a revolutionary militia group *because, despite overwhelming evidence to the contrary, he irrationally or ignorantly believes doing so is in his own best interests.* He suffers many hardships during the war. When the war ends, and because of his many acts of heroism, he becomes the leader of his country and his life goes very well for him. His life goes so well that it counterbalances the harms he faced during the war.

> *Selfless, Well-Informed Fidel:* Fidel joins a revolutionary militia group *because he wants to oust a brutal and unjust government for the sake of the oppressed*. *Due to a reliable network of informants, he knows that the government will be unsuccessful in quashing a revolution. He knows leading the revolution will incur many hardships but, after the war ends, his life will go well for him. However, these distant benefits are not what motivates him. He would have engaged in the same course of knowledge without such foreknowledge.* He suffers many hardships during the war. When the war ends, and because of his many acts of

heroism, he becomes the leader of his country and his life goes very well for him.

His life goes so well that it counterbalances the harms he faced during the war.

If selfless, well-informed Fidel is acting self-sacrificially, it must be that motivations – not foreknowledge of benefits – determine whether acts are self-sacrificial.

Motivations alone are insufficient, however, for an act to qualify as self-sacrificial. Common sense demands *some* sort of sacrifice. Consider the following case:

> *Rodrigo the Conductor*: Rodrigo is a youthful, up-and-coming talent who is motivated to act both for his own sake and for the sake of the music. Unlike Thomas, however, for Rodrigo to act for the sake of the music is for him to accept the prestigious, highly-desirable position of conductor that Thomas has just relinquished.

That does not seem self-sacrificial. Unlike Thomas, it does not seem that Rodrigo has made any sacrifices. Consider also:

> *Miguel the Politician*: Miguel, a successful businessman, is mostly motivated to run for president in order to help is country. But he is also partly motivated by his own ego. He does not endure any significant hardships along the way.

Although Miguel is mostly selfless in his motivations, he does not sacrifice anything. So he has not, intuitively, done anything self-sacrificial.

The difference between Miguel and Rodrigo on the one hand, and selfless Fidel on the other, is that Miguel and Rodrigo were motivated not only benevolently, but also selfishly. Their selfish motivations pollute their motivational set in such a way that they are not self-sacrificial.

When calculating whether an agent's act qualifies as self-sacrificial, we will be comparing some of the costs incurred to the benefits accrued. But we should not count all of the benefits (nor, indeed all of the costs). Instead we should be factoring only those costs that the agent weighed when making their decision, and only those benefits that factored into the agent's non-benevolent motives. More precisely:

> *Self-Sacrifice*$_{def}$: An agent acts self-sacrificially to the extent that 1) the agent is selflessly motivated and 2) the foreseen self-costs the agent weighed in choosing their action outweigh the benefits that non-selflessly motivated the agent.

This definition is not, prima facie, easily understood, so I will spend the next several pages explaining my use of the terms involved and applying the definition to several of this paper's examples.

First, a primer on the types of motivation. Although I do not pretend to have a satisfactory reduction of the concept of motivation to other, more fundamental concepts, I do think that motivations can usefully and intuitively be divided into two categories: selfless and non-selfless. An action is paradigmatically selflessly motivated to the extent that it is performed for the sake of something other than the agent – i.e. to act on the (implicit) assumption that something other than oneself is a final end (rather than a means). A paradigmatic way to act for the sake of something other than oneself is to act so as to benefit that thing. Alice acts for the sake of the soup kitchen patrons (i.e. she acts so as to benefit them); Thomas acts for the sake of music (i.e. he acts so as to benefit music). Let us call acting so as to benefit something other than oneself "acting from benevolent motives."

As a clarificatory aside: Is there a difference between selfless motivation and benevolent motivation? Pre-theoretically, the latter is a subcategory of the former, referring to motivations that aim at *benefitting*. So the question hinges on whether one could act for the sake of something other than oneself (i.e. act selflessly) without also acting with the aim of benefitting something other than oneself. I happen to think there is no substantive distinction between the two. But some people may believe that acting for the sake of honesty (but not for the sake of one's interlocutor), for the sake of friendship (but not for the sake of one's friend), for the sake of love (but not for the sake of any particular beloved), for the sake of charity (but not for the sake of the charity's beneficiaries), for the sake of courage, etc. can each count as selflessly motivated actions. (See the example of George and the Cherry tree, in section 4.2, above for an example of this type.)

Next, let us apply my definition of self-sacrifice to several examples for illustrative purposes. I will start with an application to *Alice's Friday Night*. Alice has two relevant options to choose from while planning out her Friday night. One option is to go to the disco with her friends. I assume the motives for this action are mixed: to some extent, Alice is motivated for her friend's sake (her friends enjoy her company); to some extent Alice is motivated for her own sake (she knows that dancing with her friends will make her happy, she knows that being happy benefits her[73], and she is motivated by this happiness-benefit). The other option is to volunteer at the soup kitchen. Alice's motives here may be mixed as well; to some extent she acts for her own sake (she takes pleasure in being of service to others) and to some extent she acts for the sake of others (the soup kitchen's hungry patrons). The following table summarizes the categorizing of the benefits to Alice, as summarized in this paragraph, and adds in some more-or-less arbitrary utility values:

[73] A subjectivist will differ from an objectivist regarding why happiness benefits Alice here, but I am not aware of any wellbeing theorist who would disputes *that* happiness benefits her.

	Disco Dancing		Soup Kitchen Volunteering	
	Benefit Description	Utility	Benefit Description	Utility
Self-benefits that factor into Alice's *self-interested* motives	Pleasure in spending time with friends	20	Pleasure in benefitting soup kitchen patrons	10
Self-benefits that factor into Alice's *selfless* motives	Satisfies preference to make friends happy	15	Satisfies preference to feed hungry soup kitchen patrons	40

What my definition of self-sacrifice instructs us to do is disregard the bottom row of this chart. To be clear, the benefits to Alice in the bottom row of the chart are real; they are just not relevant to determining whether a person's act qualifies as a sacrifice for the purpose of measuring whether an act is self-sacrificial. If we do disregard the bottom row, as suggested, then we can see that Alice's decision incurs a global opportunity cost to Alice.

It is also possible that Alice is not motivated by the pleasure she will get from volunteering at the soup kitchen. In that case, the 10 utility in the top right of the chart can also be disregarded and the act is even more self-sacrificial. (Recall the tennis player who knows of the fame and glory he will get from winning the tennis match but is simply not motivated by those consequences.)

Next I apply the definition to *Thomas the Conductor*:

	Continue Conducting		Retire	
	Benefit Description	Utility	Benefit Description	Utility
Self-benefits that factor into Thomas's non-selfless motives	Thomas enjoys being NY Symphony conductor	20	Life of retirement	10
Self-benefits that factor into Thomas's selfless motives	N/A	0	Satisfies preference that the music be as good as possible	15

Thomas benefits more from continuing to conduct (10 +15 = 25) than he does from retiring (20). But for the purpose of determining if Thomas's decision to retire is self-sacrificial, we disregard the bottom row, i.e. those benefits to Thomas that factored into his benevolent motives. Once we have made that discount, Thomas's decision to retire has a 10 utility (global) opportunity cost (20 – 10 = 10).

And finally I apply my definition of self-sacrifice to the case of Mamoudou:

	Continue Walking		Attempt to Rescue Child	
	Benefit Description	Utility	Benefit Description	Utility
Self-benefits that factor into Mamoudou's non-selfless motives	No risk of injury (from falling multiple stories)	15	Pleasure in helping others; Praise for being a hero	5
Self-benefits that factor into Mamoudou's selfless motives	Default	0	Satisfies preference to save child	20

One might wonder here why the award of French citizenship to Mamoudou is not included in the chart. The answer is simply that my definition of self-sacrifice includes only those predicted benefits that factored in to Mamoudou's motivations. Recall that when determining if clause 2 of my definition of self-sacrifice is satisfied only those benefits that the actor weighed are relevant. Mamoudou could not have been influenced by the future benefit of French citizenship because he was not in a position (epistemically) to predict that consequence.

Earlier in this section, I set aside the debate over whether there were any selfless motives that were not also benevolent motives. Here I explore a somewhat related objection regarding borderline cases of selflessness or non-selflessness. The following example is due to Richard Miller:[74]

> *Richard's Book*: Suppose Richard publishes a book and wants it to do well (i.e. become widely read, be influential, etc.). So Richard engages in some standard book-promotion efforts. These efforts take up time and energy that could have been spent on more enjoyable leisure activities. Note that Richard would not have been motivated to engage in book-promotion efforts if this same book had been written by someone else.

Is Richard acting self-sacrificially? The answer to that question straightforwardly depends on the (non-straightforward) question of whether Richard is selflessly motivated. Could acting for the success of *Richard's own book* count as selfless? I have two answers to offer here. First, I suspect many people will have unclear intuitions regarding whether Richard is acting self-sacrificially in the first place. If that is the appropriate reaction to the case, then the fact that it is unclear whether Richard acts selflessly – and therefore also unclear whether my definition judges that he acts self-sacrificially – means that my definition does not face an extensional problem.

[74] In conversation.

Second, I will argue that Richard *is* acting selflessly. My argument relies on a distinction between treating something as a means for self-interested reasons and treating something as a final end for selfish / self-guided reasons. Consider the following analog to Richard's Book:

> *Darth Vader's Son*: Darth Vader risks his life (and career) for the sake of his son (by attacking Emperor Palpatine). However, Darth Vader would not have taken these risks for someone else's son (all else equal).[75]

Here I think it is intuitive that Vader has acted selflessly and self-sacrificially. Albeit, Vader was, in a sense, selfish (or at least self-guided) in picking out *who* to treat as a final end. But *why* an agent treats some particular other thing as a *final* end is irrelevant to the question of whether the agent acted selfishly, as *Darth Vader's Son* illustrates. Applying this analysis to *Richard's Book*, it becomes clear that although it was, to a degree, selfish of Richard to *pick his own book* as worthy of being a final end, it is still possible for him to act selflessly for the sake of that end. To pump this intuition larger, suppose the publisher published the book under a pseudonym. If this has no impact on Richard's promotion efforts, then we would have some evidence that Richard really was treating the book as a final end rather than as a means (e.g. for his own standing among his peers or the public). (Though of course if Richard wanted the book to succeed merely as a means for his self-esteem, that would be self-interested since he would not, then, be treating the success of the book as a *final* end.)

To reiterate: my analysis of self-sacrifice is independent of any particular theory of wellbeing. I am also not proposing a change in how we *ought* to use the concept of self-sacrifice. Instead, I am proposing that self-sacrifice, although intuitively grasped by ordinary speakers, turns out to have a somewhat complicated formula.

The upshot of my proposal is that the problem of self-sacrifice for wellbeing subjectivism turns out to have been founded on a misunderstanding of what self-sacrifice required. Actions that we pretheoretically thought were self-sacrificial *are* self-sacrificial, even if wellbeing subjectivism is true.

[75] The motivations are analogous, not the characters!

Works Cited

Bradley, Ben. 2016. “Well-Being at a Time.” *Philosophic Exchange* 45 (1). https://digitalcommons.brockport.edu/phil_ex/vol45/iss1/1.

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

Campbell, Stephen M. 2013. “An Analysis of Prudential Value.” *Utilitas* 25 (3): 334–54.

Carson, Thomas L. 2000. *Value and the Good Life*. University of Notre Dame Press.

Darwall, Stephen L. 2002. *Welfare and Rational Care*. Princeton, N.J.: Princeton University Press. https://ebookcentral.proquest.com/lib/cornell/detail.action?docID=557139.

Daveney, T. F. 1961. “Wanting.” *Philosophical Quarterly* 11 (April): 135–44.

Foot, Philippa. 1972. “Morality as a System of Hypothetical Imperatives.” *Philosophical Review* 81 (3): 305–16.

Gardner, Molly. forthcoming. “What Is Harming?” In *Principles and Persons: The Legacy of Derek Parfit*. Oxford University Press.

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

Heathwood, Chris. 2005. “The Problem of Defective Desires.” *Australasian Journal of Philosophy* 83 (4): 487–504.

———. 2011. “Preferentism and Self-Sacrifice.” *Pacific Philosophical Quarterly* 92 (1): 18–38.

———. 2017. “Which Desires Are Relevant to Well-Being?” *Noûs*. https://doi.org/10.1111/nous.12232.

Hume, David. 1739. *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning Into Moral Subjects*. Oxford University Press.

Lewis, David. 1988. “Desire as Belief.” *Mind* 97 (418): 323–32.

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

———. 2019. “Why Subjectivists About Welfare Needn’t Idealize.” *Pacific Philosophical Quarterly* 100 (1): 2–23. https://doi.org/10.1111/papq.12232.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Oxford Clarendon Press.

Overvold, Mark Carl. 1980. “Self-Interest and the Concept of Self-Sacrifice.” *Canadian Journal of Philosophy* 10 (1): 105–18. https://doi.org/10.2307/40231134.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford University Press.

———. 2011. *On What Matters*. Oxford University Press.

Rosati, Connie S. 2009. “Self-Interest and Self-Sacrifice.” *Proceedings of the Aristotelian Society* 109 (1 pt. 3): 311–25.

Schapiro, Tamar. 2014. “What Are Theories of Desire Theories Of?” *Analytic Philosophy* 55 (2): 131–50.

Schueler, G. F. 1995. *Desire: Its Role in Practical Reason and the Explanation of Action*. Cambridge, Mass.: MIT Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1723.

Sobel, David. 2016. *From Valuing to Value: A Defense of Subjectivism*. Oxford University Press.

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

Williams, Bernard. 1981. *Moral Luck: Philosophical Papers, 1973-1980*. Cambridge [Cambridgeshire]: Cambridge University Press.

Williams, Bernard A. O. 1973. *Problems of the Self*. Cambridge University Press.