JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXIX · NUMBER 2
*January 2025*

ARTICLES



DISCUSSIONS



OPEN ACCESS FREE FOR AUTHORS AND READERS ISSN 1559-3061

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXIX · NUMBER 2
*January 2025*

ARTICLES



DISCUSSIONS

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
*http://www.jesp.org*
ISSN 1559-3061

The *Journal of Ethics and Social Philosophy* (*JESP*) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge. Articles are typically published under the CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 license, though authors can request a different Creative Commons license if one is required for funding purposes. Funding for the journal has been made possible through the generous commitment of the Division of Arts and Humanities at New York University Abu Dhabi.

*JESP* aspires to be the leading venue for the best new work in the fields that it covers, and it is governed by a correspondingly high editorial standard. The journal welcomes submissions of articles in any of these and related fields of research. The journal is interested in work in the history of ethics that bears directly on topics of contemporary interest, but does not consider articles of purely historical interest. It is the view of the editors that the journal's high standard does not preclude publishing work that is critical in nature, provided that it is constructive, well argued, current, and of sufficiently general interest. *JESP* also endorses and abides by the Barcelona Principles for a Global Inclusive Philosophy, which seek to address the structural inequality between native and nonnative English speakers in academic philosophy.

*JESP* publishes articles, discussion notes, and occasional symposia. Articles normally do not exceed 12,000 words (including notes and references). *JESP* sometimes publishes longer articles, but submissions over 12,000 words are evaluated according to a proportionally higher standard. Discussion notes, which need not engage with work that was published in *JESP*, should not exceed 3,000 words (including notes and references). *JESP* does not publish book reviews.

Papers are published in PDF format at https://www.jesp.org. All published papers receive a permanent DOI and are archived both internally and externally.

*Journal of Ethics and Social Philosophy*
VOL. 29, NO. 2 · JANUARY 2025
*https://doi.org/10.26556/jesp.v29i2.3558*


# THE OVERWEIGHTED INTEGRITY PROBLEM

## CONSCIENCE, COMPLICITY, AND MORAL STANDING

*Kyle G. Fritz*

When she was eighteen weeks pregnant, Tamesha Means suffered a ruptured amniotic sac. The hospital where she presented, the only one in her county in Michigan, was Catholic. At eighteen weeks, the fetus was not viable, and an abortion would have been the safest option. Nevertheless, Means was given two Tylenol and sent home. She presented two more times, bleeding and in severe pain, but it was only when she went into labor that the hospital provided care. The baby died within hours.[1]

The health care professionals at the hospital did not tell Means that her fetus would not survive or that an abortion could reduce serious health complications for her. In fact, Means had an infection of the fetal membranes and umbilical cord as a result of the amniotic rupture.[2] While one might think that health care professionals and institutions are legally required to disclose medically relevant information to patients, the hospital was protected from malpractice claims because conscience law in Michigan requires only that providers disclose "morally legitimate alternatives" to the recommended treatment. Since the hospital was Catholic and did not see abortion as a morally legitimate alternative, they were not required to disclose that option.[3]

It is unsurprising that a Catholic hospital would refuse to *perform* an abortion for Means. Many states in the United States have conscience laws protecting conscientious refusal to perform some medical service. For instance, in Mississippi, "a health-care provider may decline to comply with an individual instruction or health-care decision for reasons of conscience."[4] What is striking,

1 Kaye et al., *Health Care Denied*, 9–10.

2 Sawicki, "The Conscience Defense to Malpractice," 1257–58.

3 Sawicki, "The Conscience Defense to Malpractice," 1259–60. Means is not a one-off case. See similar cases detailed in Kaye et al., *Health Care Denied*; and National Women's Law Center, "Below the Radar."

4 Miss. Code § 41-41-215(5). Conscientious refusals are also protected federally through the Church Amendments (42 U.S.C. § 300a-7 et seq.), the Public Health Service Act (42 U.S.C. § 238n), the Weldon Amendment, and the Affordable Care Act (§ 1303(b)(4)).

however, is that the hospital was not even legally required to provide Means with all the medically relevant *information* about her situation so that she could make an informed decision about her health. By not providing information, the professionals at the hospital could avoid any complicity in perceived wrongdoing if Means chose to travel outside the county to seek an abortion.

Providing information is not the only way in which someone may believe they are complicit in wrongdoing. A part-time admissions clerk refused to type lab and admissions forms for abortion patients, while another employee refused to clean surgical tools used in abortion.[5] Depending on the state, some of these objections too might be protected under conscience law. Title 16, section 51.41 of the Pennsylvania Administrative Code protects those who object to even "cooperating in abortion or sterilization," where such cooperation can include "disposal of or assistance in the disposal of aborted fetuses" and "cleaning the instruments used in the abortion or sterilization procedure."[6]

In fact, many state conscience laws protect health care professionals and providers from being even indirectly involved with some procedure they find objectionable. In her excellent study of state conscience laws, Nadia Sawicki finds that of the states that protect a right to refuse to participate in abortion, only Illinois requires that providers inform patients of all available treatment options, including abortion.[7] In most states, providers are not required to disclose to patients that abortion may be medically appropriate and available elsewhere.[8]

Of course, these policies are not restricted to abortion. Mississippi's Health Care Rights of Conscience Act is perhaps the broadest example, granting health care providers the right to conscientiously not participate in "any phase of patient medical care, treatment or procedure, including, but not limited to, the following: patient referral, counseling, therapy, testing, diagnosis or prognosis, research, instruction, prescribing, dispensing or administering any device, drug, or medication, surgery, or any other care or treatment rendered by health care providers or health care institutions."[9] Notably, the law covers all types of health care professionals, and it is increasingly common to find legal protections for not only physicians but also nurses, pharmacists, emergency medical technicians, physician assistants, public health officials, medical

5 Pope, "Conscience Clauses and Conscientious Refusal," 165.

6 PA Code § 16.51.41. Sawicki clarifies that Pennsylvania does not include recordkeeping in its understanding of cooperation, and so refusing to type lab forms likely would not be protected under Pennsylvania law ("The Conscience Defense to Malpractice," 1265n47).

7 Sawicki, "The Conscience Defense to Malpractice," 1285.

8 Sawicki, "The Conscience Defense to Malpractice," 1283.

9 Miss. Code Ann. § 41-107-3a.

students, researchers, and even institutional health care providers like hospitals and skilled nursing facilities.[10]

In a similar vein, Oklahoma's Freedom of Conscience Act allows health care professionals to refuse to "perform, practice, engage in, assist in, recommend, counsel in favor of, make referrals for, prescribe, dispense, or administer drugs or devices or otherwise promote or encourage" certain health care services, including abortion, reproductive assistance technology, and medical aid in dying (MAID).[11] Other states have attempted to pass similar legislation, with varying success.[12] Crucially, few of these conscience clauses include exceptions for emergency situations.[13]

In this paper, I argue that conscience policies that seek to protect health care professionals from any kind of association with medically accepted care to which they object are unjust. Such policies are often defended because they protect the *integrity* of health care professionals. While this is admittedly important, these policies nevertheless grant too much weight to that integrity in light of competing patient interests and values. Despite the significant attention given to conscientious refusal to *perform* some service, as well as to the duty of referral and whether individuals are *actually* complicit in some activity, too little attention has been given to just how wide-ranging many conscience policies currently are and why these policies are unjust.

I begin in section 1 by explaining the connection between conscience and integrity and the value of integrity. In section 2, I argue that despite its value, protecting integrity even in these indirect cases of complicity requires compromising other key values like autonomy and leads to significant harms. Accordingly, these policies overweight integrity and are unjust. In section 3, I explore whether other considerations in addition to integrity might shift the balance in favor of these policies. I deny that tolerance will provide the needed additional weight, but one unique proposal is the interest the state has in protecting the moral standing of its citizens to hold each other accountable. Despite its initial promise, I argue in section 4 that unwillingly complicit professionals do not necessarily lose their standing, so this cannot serve as an additional weighty consideration for these policies. Consequently, I conclude that such policies are unjustified and should be restricted.

10 Sawicki, "The Conscience Defense to Malpractice," 1263–64.

11 Oklahoma Code § 63-1-728.

12 Pope, "Conscience Clauses and Conscientious Refusal"; and Sawicki, "The Conscience Defense to Malpractice."

13 Wicclair, *Conscientious Objection in Health Care*, 211.

## 1. THE SIGNIFICANCE OF INTEGRITY

There are various reasons to allow for conscientious refusal in our policies. These policies promote diversity and tolerance and encourage those who are ethically sensitive to join the medical profession.[14] Yet as Mark Wicclair writes, "moral integrity is among the most frequently cited reasons for accommodation—both by its defenders and its critics."[15] Indeed, many in the debate see it as the *strongest* reason for accommodating such refusal.[16] I agree, and I will accordingly focus on integrity in this paper.[17]

Integrity seems the strongest reason for protecting conscientious refusal in part because of the nature of conscience and the value of integrity. Conscience tracks one's moral integrity by advocating for one's deeply held core commitments. Maintaining moral integrity requires acting in accordance with these core commitments and judgments. Sometimes these judgments will conflict with current medical practices, so conscientious refusal allows someone to preserve their integrity.[18] Because integrity is quite valuable, there is good reason to protect it in our policies if we can.

14 Wear, Lagaipa, and Logue, "Toleration of Moral Diversity and the Conscientious Refusal by Physicians to Withdraw Life-Sustaining Treatment"; and Wicclair, *Conscientious Objection in Health Care*, 29. I revisit tolerance in section 3 below.

15 Wicclair, *Conscientious Objection in Medicine*, 8.

16 Wester, "Conscientious Objection by Health Care Professionals," 429. See also Benn, "Conscience and Health Care Ethics"; Birchley, "A Clear Case for Conscience in Healthcare Practice"; Brock, "Conscientious Refusal by Physicians and Pharmacists"; Magelssen, "When Should Conscientious Objection Be Accepted?"; and Wicclair, *Conscientious Objection in Health Care*.

17 Those who do not find integrity a plausible justification for conscientious refusal can see my argument conditionally: if integrity *were* to justify conscientious refusal, it would not do so to the extent currently enshrined in policy.

18 Ben-Moshe argues that allowing for conscientious refusal protects individuals from doing not merely what they *believe* to be wrong but what is *actually wrong*. Drawing inspiration from Adam Smith, Ben-Moshe argues that when a health care professional reasons from the standpoint of an impartial spectator and consults their conscience, "[their] claims of conscience are true, or at least approximate moral truth to the greatest degree possible for creatures like us, and should thus be respected" ("The Truth Behind Conscientious Objection in Medicine," 404). I set this view aside for two key reasons. First, it relies on a controversial and, to my mind, implausible view of ethical justification and truth. Second, I am unconvinced that it is sufficiently clear what an impartial spectator might judge in contentious medical cases such as first-trimester abortions and MAID. See Wicclair, *Conscientious Objection in Medicine*, 53–54. Addressing this view fully is outside the scope of this paper, however, so I assume here that protecting judgments of conscience is important not because it tracks moral truth but rather because it protects one's integrity.

Some, like Wicclair, see integrity as *intrinsically* valuable.[19] That is, having core moral commitments and being disposed to act on them is valuable in itself; all else equal, the world is a better place if it includes such individuals of integrity rather than people who act only opportunistically or transactionally.[20]

Even if integrity is not intrinsically valuable, it may nevertheless be quite valuable. Integrity preserves our dignity.[21] It also allows individuals to remain authentic to their true selves. Because the moral judgments associated with one's conscience are often a crucial part of who one is, they bear on one's self-conception or identity.[22] Being forced to act against those beliefs can consequently seem like an act of self-betrayal, which leads to a loss of self-respect.[23] Indeed, some have claimed that "a basic part of an acceptable human life is to live in accordance with one's deeply held beliefs and values."[24] Ben-Moshe goes so far as to suggest that "sometimes life might not be worth living if it cannot be lived according to [one's] evaluative judgments."[25]

Clearly, integrity can be compromised when one is forced to participate in an activity to which one objects. Physicians who are forced to provide MAID when they believe killing is wrong may feel that they have lost a crucial part of their identity in the process. Yet some champions of conscientious refusal have also held that integrity can be compromised even through more indirect involvement. Karen Brauer, president of Pharmacists for Life, explained her objection to referring an individual or providing them information this way: "That's like saying, 'I don't kill people myself but let me tell you about the guy down the street who does.' What's that saying? 'I will not off your husband, but I know a buddy who will?' It's the same thing."[26] The complaint here is that even if someone does not directly participate in the activity, if they are indirectly

19 Wicclair, *Conscientious Objection in Health Care*, 27.

20 Wicclair, *Conscientious Objection in Medicine*, 9.

21 Dworkin, *Life's Dominion*, 229–30.

22 Brock, "Conscientious Refusal by Physicians and Pharmacists," 189.

23 Ben-Moshe, "The Truth Behind Conscientious Objection in Medicine," 404; Blustein, "Doing What the Patient Orders," 296. Some have argued that even *institutions* that are forced to go against the mission and values that comprise their identity may be said to lose integrity. See Wicclair, *Conscientious Objection in Health Care*, 148–52; and Sulmasy, "What Is Conscience and Why Is Respect for It So Important?" 142–44.

24 Brudney and Lantos, "Agency and Authenticity," 223.

25 Ben-Moshe, "Internal and External Paternalism," 676.

26 Cited in Shahvisi, "Conscientious Objection," 84.

involved in it in some way, they thereby become complicit in the alleged wrongdoing, and their integrity is compromised due to that complicity.[27]

Notably, a health care professional may feel complicit in perceived wrongdoing in a variety of ways. As indicated in Brauer's quote, referrals are one such way. In fact, some have argued that an objection to performing some service already implicitly involves an objection to referring for that service.[28] As Card writes, "it is unclear what actual ethical difference exists" between the duty to refer and the duty to directly provide the service, precisely because one remains part of this causal chain of events.[29]

A health care professional can merely provide information about possible services and procedures without making a referral to another provider. Yet one may worry that this too makes one complicit in wrongdoing, since without that information the patient might not seek out the service elsewhere on their own. Other examples of health care professionals raising concerns about complicity include an emergency medical technician refusing to drive a woman suffering from abdominal pain to an abortion clinic and a county health department employee refusing to translate information on family planning and abortion options into Spanish.[30] In each of these cases, even though the individual is not themself performing the procedure to which they object, they are nevertheless somehow involved in the procedure. This involvement may cause them to feel complicit in perceived wrongdoing, threatening their integrity and perhaps even their self-identity. Call such refusal to be involved in any way with a procedure to which one objects *complicity refusal*.

Some have suggested that while health care professionals may be complicit in these cases, complicity comes in degrees, and referring and informing are minimal forms of complicity that are not morally problematic.[31] Yet drawing the line between what degree of complicity is acceptable itself requires contentious ethical judgments.[32] Because how much complicity is permissible is a matter for conscience as well, individuals will differ in where they draw the

27 Bayles, "A Problem of Clean Hands," 167; Clarke, "Conscientious Objection in Healthcare, Referral and the Military Analogy," 220; Shahvisi, "Conscientious Objection"; and Ben-Moshe, "Conscientious Objection in Medicine."

28 Hill, "Abortion and Conscientious Objection," 347.

29 Card, "Conscientious Objection and Emergency Contraception," 9.

30 Pope, "Conscience Clauses and Conscientious Refusal,"165.

31 Brock, "Conscientious Refusal by Physicians and Pharmacists," 197. See also Sulmasy, "What Is Conscience and Why Is Respect for It So Important?" 141.

32 Wicclair, *Conscientious Objection in Health Care*, 41.

line.[33] For some, "even a minimal degree of complicity would represent a serious violation of their moral integrity."[34]

It is this feeling of complicity or belief that one's integrity has been compromised that is important. Integrity is a *subjective* matter.[35] Even if abortion is entirely morally innocuous, someone who nevertheless believes it is tantamount to murder will still believe that their integrity is compromised if they are somehow associated with the procedure. Whether abortion is actually wrong and whether the individual is actually blameworthy for wrongdoing are irrelevant to their beliefs and their felt integrity violation. They will still feel the telltale pangs of guilt associated with tarnished integrity.[36]

Given the significance of integrity as well as the facts that it can be compromised when one must act against one's conscience and that one's conscience may demand that one not be involved in the perceived wrongdoing in any way, current policies protecting complicity refusals like those I surveyed above may seem justified. In the next section, however, I argue that this justification is merely apparent.

33 Sulmasy, "What Is Conscience and Why Is Respect for It So Important?" 142.

34 Minerva, "Conscientious Objection, Complicity in Wrongdoing, and a Not-So-Moderate Approach," 118. Similarly, Blustein suggests that even if one believes some service is generally wrong, informing or referring the patient in certain circumstances is, all things considered, morally permissible ("Doing What the Patient Orders," 314). But again, the objector may not always make this judgment, and in fact many of them do not.

35 Gerrard, "Is It Ethical for a General Practitioner to Claim a Conscientious Objection When Asked to Refer for Abortion?" 600; Sepinwall, "Conscientious Objection, Complicity, and Accommodation," 206; and Wicclair, "Conscientious Objection in Healthcare and Moral Integrity," 12.

36 One might respond that if an individual feels so violated, they simply ought to leave the profession, or at least shift to a subfield compatible with the requirements of their conscience. See Brock, "Conscientious Refusal by Physicians and Pharmacists"; and Stahl and Emanuel, "Physicians, Not Conscripts." After all, entering a profession is a voluntary choice, and objectors should have known that their job would involve actions that could conflict with their conscience and threaten their integrity. See, e.g., Schuklenk, "Conscientious Objection in Medicine." While I do not disagree, we must appreciate that health care professionals *do* know what they are getting into: a field that explicitly allows for conscientious refusal. See Robinson, "Voluntarily Chosen Roles and Conscientious Objection in Health Care," 721. In many states, policies protecting complicity refusal are already in place, so someone entering the field could reasonably expect that their right to refuse even indirect involvement in some perceived wrongdoing would be legally protected. This only bolsters my point that legal protection of complicity refusal is too broad, because refusing to inform or refer patients clearly conflicts with professional obligations to care and advocate for patients and promote their health and well-being, and the law should better reflect the professional obligations of medical professionals. Thanks to two anonymous referees for encouraging me to address this point.

## 2. THE OVERWEIGHTED INTEGRITY PROBLEM

Whether it is intrinsically valuable or merely instrumentally valuable, integrity provides only a *pro tanto* reason for protecting conscience.[37] We also must consider competing reasons and values, including what is threatened or lost when we protect integrity in our policies. Accommodating someone's refusal to perform a medical service can be burdensome. It is burdensome for patients who will face delays in receiving care while they wait for a willing professional. It is burdensome for other professionals, who must take on additional work. Nevertheless, if these burdens are acceptably small, it may be justified to protect integrity.[38]

There are various ways to keep these burdens minimal when a professional refuses to perform some service: ensuring that there are enough willing providers within a certain geographical area, careful management of staff, etc.[39] Accordingly, it may be reasonable to protect conscientious refusal to directly provide some service in such cases. Yet I am focused not on *direct conscientious refusal* but rather on complicity refusal. It is much more difficult to keep someone from being involved *in any way* with procedures to which they object while keeping the burdens to patients and coworkers at acceptable levels.

Protecting complicity refusal can have serious consequences for patients. The types of treatments institutions and individuals typically object to are concerned with beginning- or end-of-life care and can be life altering.[40] For instance, in ectopic pregnancies, the fertilized egg implants and grows outside of the uterus, which means the developing embryo cannot survive. If left untreated, the embryo can cause serious harm to surrounding organs and lead to the death of the mother.[41] Although one treatment option is the termination of the pregnancy, many Catholic institutions will not even inform the patient of that option, let alone assist in referral. When patients are not informed of key options, including termination of pregnancy, their lives and well-being are put at serious risk.[42]

Of course, not every pregnancy will be life-threatening in this way. But withholding information about options regarding abortion can still lead to a delay in the actions that a patient takes, limiting their family planning options. Some

37 Wicclair, *Conscientious Objection in Medicine*, 8.

38 Magelssen, "When Should Conscientious Objection Be Accepted?" 19.

39 Minerva, "Conscientious Objection, Complicity in Wrongdoing, and a Not-So-Moderate Approach," 116.

40 Ben-Moshe, "The Truth Behind Conscientious Objection in Medicine," 405.

41 National Women's Law Center, "Below the Radar," 4–5.

42 Kaye et al., *Health Care Denied*; and Uttley et al., "Miscarriage of Medicine."

states have stricter laws regarding second- and third-trimester abortions, which means that if patients do not learn of care options early or if they have their care significantly delayed due to someone's complicity refusal, they may have little option but to complete the pregnancy.[43] This can have a drastic impact on the mother's life as well as the child's.

These consequences illustrate the way in which complicity refusal has the potential to violate several key values, some of which are familiar in biomedical ethics. Perhaps most salient is patient *autonomy*, a crucial value of self-direction regarding one's life.[44] Patients who are not informed of all the medically relevant options cannot make informed decisions about their own health. This was illustrated in the case of Tamesha Means, though refusals to translate information also run afoul of patient autonomy. Without autonomy, it can be hard for patients to live their lives authentically in the way they want. Indeed, just as one may feel an acceptable life requires the ability to live with integrity, an acceptable life plausibly requires a high degree of autonomy.

Additionally, the principle of *beneficence* values enhancing the welfare of others, and the principle of *nonmaleficence* calls for avoiding imposing harm on others.[45] Both of these principles are threatened by complicity refusal. As we have seen, those who are unaware that abortion could save their life or protect their health are at significant risk of physical and psychological harm. This is, of course, to say nothing of the professional obligations a health care professional has to their patients and to ensuring they are cared for.[46]

Refusing to refer a patient for certain kinds of reproductive care or emergency contraception may also reinforce an oppressive social norm that can increase a patient's feeling of social stigma.[47] When patients feel vilified, their *moral identity* as a good person and sense of self-respect may consequently be threatened.

Integrity is an important value. But we must consider the consequences of protecting integrity to the extent we do in complicity refusal, as well as the way moral and professional values like autonomy, beneficence, and nonmaleficence are threatened. Integrity may protect one's moral identity and self-respect, but protecting complicity refusal may sometimes threaten the moral identity and

43 Sawicki, "Mandating Disclosure of Conscience-Based Limitations on Medical Practice," 97.

44 Beauchamp and Childress, *Principles of Biomedical Ethics*.

45 Beauchamp and Childress, *Principles of Biomedical Ethics*.

46 Brock, "Conscientious Refusal by Physicians and Pharmacists," 192; May and Aulisio, "Personal Morality and Professional Obligations," 32; and Sawicki, "The Conscience Defense to Malpractice," 1295–301.

47 McLeod, *Conscience in Reproductive Health Care*, 52–55. This is not to say the professional themself endorses oppression or intends to ostracize patients, but this may be an unwelcome byproduct of refusal.

self-respect of patients. Integrity alone does not seem weighty enough to compete with these other values and consequences. This is plausibly true even if integrity is intrinsically valuable, as autonomy also has a strong claim to intrinsic value.

The case is even stronger when one considers that integrity is plausibly violated to a lesser degree in cases of complicity. If someone is forced to kill despite moral objections, they are *the cause* of an individual's death. Yet if someone is complicit in such an act, they are merely a part of the causal chain that leads to the individual's death. This is not to say the complicit individual will not feel their integrity has been violated, but directly participating in an act that they believe is wrong is likely a greater affront to integrity than complicity. That this is so is borne out upon reflection: if forced to choose between performing an act one thinks is wrong and being complicit in such an act, I suspect nearly everyone would prefer complicity. While integrity is important, then, it plausibly bears less weight in indirect cases of complicity.[48]

To summarize, the value of integrity may plausibly justify protecting direct conscientious refusal, provided steps are taken to ensure burdens to patients are minimized. But I contest that it is not valuable enough to outweigh the significant harms, burdens, and value violations that result from protecting that integrity from all possible violations. Current policies that protect health care professionals from even indirect involvement in a procedure to which they object overweight the professionals' integrity compared to the interests of patients and competing values. I call this the *overweighted integrity problem*. The overweighted integrity problem is, in my view, a powerful reason why conscience clauses, at least in medicine, should be worded in a more restrictive way

48 An anonymous referee worries that we cannot compare integrity with other values or reasonably weigh integrity against other consequences; perhaps these are simply incommensurable values. While I admit that these values may *seem* incommensurable, I am also partial to Schmidtz's insight: "At some level, commensuration is always *possible*, but there are times when something (our innocence, perhaps) is lost in the process of commensurating" ("Value in Nature," 394). Integrity is indeed valuable. Yet so are autonomy, health, and well-being. "The hard fact is that priceless values can come into conflict. When they do, and when we rationally weigh our options, we put a price, in effect, on something priceless. . . . The world hands us painful choices. Weighing our options is how we cope" ("Value in Nature," 393). Nevertheless, even if these values are incommensurable, that does not entail they are *incomparable*. See Chang, "Value Incomparability and Incommensurability": items are incommensurable when "there is no cardinal unit of measure that can represent the value of both items" (207), but they are incomparable when "they fail to stand in an evaluative comparative relation" (205). In that case, my talk of scales and weighing may be inapt metaphorically, but even if these values are incommensurable, they are nevertheless comparable, and protecting autonomy, health, and well-being is more important than protecting integrity in the case of complicity refusal.

to rule out complicity refusal. While integrity is valuable, competing values and consequences are weightier.

Advocates of complicity refusal might initially resist this conclusion in several ways. First, despite the high stakes for patients, if complicity refusal is uncommon, these negative consequences and value violations may occur only rarely, leaving a stronger case for protecting integrity. It is difficult to know exactly how many patients might be impacted by conscientious refusal, let alone how many are impacted by complicity refusal, since many of these refusals will go undocumented. Nevertheless, there are some data that can be useful, especially regarding reproductive care. Conscientious objection to reproductive services is common at the institutional level and may affect millions of patients.[49] Four of the ten largest US hospital systems are Catholic, and Catholic hospitals treat one out of every seven patients, yet they almost universally refuse to provide abortions or sterilizations.[50] Presumably, such refusals spill over into mere complicity refusals, as was the case with Tamesha Means.

Although individual conscientious refusal is not as common as institutional refusal, survey studies of health care professionals suggest a sizeable portion value their integrity even at the expense of patient autonomy.[51] For instance, 22 percent of US primary care physicians surveyed disagreed with the statement "Physicians should not let their religious beliefs keep them from providing patients legal medical options."[52] Similarly, in a survey of over one thousand Idaho nurses, almost 25 percent responded that a nurse's right to conscientious objection should take precedence over a patient's right to health care choices.[53] And in a survey of gynecologic oncologists, 45 percent of those surveyed reported that their personal religious and spiritual beliefs "play a role in the medical options they offered patients."[54] The takeaway lesson is that complicity refusals from both institutions and individual health care professionals may be more common than many of us realize:

> If physicians' ideas translate into their practices, then 14% of patients—more than 40 million Americans—may be cared for by physicians

49 Sawicki, "The Conscience Defense to Malpractice," 1287.

50 Sawicki, "The Conscience Defense to Malpractice," 1288.

51 Sawicki, "The Conscience Defense to Malpractice," 1290.

52 Lawrence and Curlin, "Autonomy, Religion and Clinical Decisions," 216.

53 Davis, Schrade, and Belcheir, "Influencers of Ethical Beliefs and the Impact on Moral Distress and Conscientious Objection," 745.

54 Ramondetta et al., "Religious and Spiritual Beliefs of Gynecologic Oncologists May Influence Medical Decision Making," 576. It should be noted that the response rate for the survey was 14 percent, and Ramondetta and colleagues recommend further research.

> who do not believe they are obligated to disclose information about medically available treatments they consider objectionable. In addition, 29% of patients—or nearly 100 million Americans—may be cared for by physicians who do not believe they have an obligation to refer the patient to another provider for such treatments.[55]

The likelihood of a patient being affected by a complicity refusal is not insubstantial.

Despite the probable frequency of complicity refusal, one might suggest that we can mitigate the harms of such refusals and thereby protect competing values more easily than I suggest. For instance, physicians could discuss with new patients at the outset that they have moral objections to certain procedures and will not perform or refer patients for such procedures nor inform them when such procedures might be medically relevant options.[56] Yet physicians cannot know beforehand all the relevant procedures that may apply to a new patient, and such a conversation at an initial meeting might be overwhelming and stressful for patients. Alternatively, physicians or institutions could post signs clearly indicating that they do not offer certain services.[57] While this may avoid some harms to patients, it is far from clear it reduces them sufficiently. Someone like Means may not even have known that abortion was a possible treatment for her condition, so knowing abortions are not offered at that hospital would not have been helpful. Respecting patient autonomy and self-determination requires ensuring patients have information about what their relevant options are, and signage does not provide this knowledge.

Similar issues arise with Ben-Moshe's creative suggestion that objectors advertise their conscientious objections in a publicly accessible online database and allow patients to choose practitioners who do not object to some practice.[58] First, as with posted signs, this still assumes that patients will have the relevant knowledge of which procedures they need in order to search the database effectively. Connecting patients with advocacy groups to help them navigate such issues requires significant resources, and it would also plausibly be a significant source of anxiety for patients. Additionally, as Ben-Moshe

55 Curlin et al., "Religion, Conscience, and Controversial Clinical Practices," 597. Notably, this data concerning referrals is lower than other researchers have reported. In a survey of two thousand US physicians, Combs et al. found that 43 percent disagreed that physicians are obligated to make referrals that they believe are immoral ("Conscientious Refusals to Refer," 399).

56 Wear, Lagaipa, and Logue, "Toleration of Moral Diversity and the Conscientious Refusal by Physicians to Withdraw Life-Sustaining Treatment," 155.

57 Minerva, "Conscientious Objection, Complicity in Wrongdoing, and a Not-So-Moderate Approach," 117; and Dresser, "Professionals, Conformity, and Conscience," 10.

58 Ben-Moshe, "Conscientious Objection in Medicine."

acknowledges, this solution is limited in that it does not work in emergencies or geographically limited areas.[59]

Although we ought to pursue routes to limit harms to patients whenever feasible, I am unconvinced that the proposals above sufficiently address these harms or the threats to autonomy and self-respect that complicity refusal also poses. When considering just how broad policies protecting conscientious refusal are, it can be quite difficult to respect an individual's wish to not be associated with some perceived wrongdoing in any way. Impactful attempts to mitigate harm to patients unfortunately come at the cost of significant resources, trading some negative consequences for others without significantly shifting the balance on the scale.

I have argued that current policies protecting complicity refusal overweight integrity in the face of competing values and harms, and consequently such policies should be reformed to better balance integrity with these other values and consequences. This is not to deny that protecting the conscience, and consequently the integrity, of individuals is an "important component . . . of our social and political structures."[60] It is rather to point out that the state has competing interests, including protecting the health and autonomy of its citizens. Even if the state can balance integrity and competing values in direct conscientious refusal, it is implausible that it can do so for complicity refusal. Nevertheless, one might insist that there are yet other considerations in addition to integrity that could justify the state legally protecting complicity refusal, and so these policies do not overweight integrity after all. I turn to these considerations next.

## 3. Integrity, Tolerance, and Moral Standing

In addition to integrity, tolerance has been offered as a reason to protect conscientious refusal. For instance, Sulmasy writes, "Respect for conscience is at the root of the concept of tolerance. I define tolerance as mutual respect for conscience."[61] Wear and colleagues argue that requiring physicians to refer for care that they find objectionable "lacks any sensitivity toward or toleration of such moral views."[62]

59 Ben-Moshe, "Conscientious Objection in Medicine," 282. Wicclair, "Commentary," offers additional concerns with Ben-Moshe's proposal. While I find these concerns compelling, I cannot devote more attention to them here. Thanks to an anonymous referee for encouraging me to consider these methods of mitigating harms.

60 May and Aulisio, "Personal Morality and Professional Obligations," 33.

61 Sulmasy, "What Is Conscience and Why Is Respect for It So Important?" 145.

62 Wear, Lagaipa, and Logue, "Toleration of Moral Diversity and the Conscientious Refusal by Physicians to Withdraw Life-Sustaining Treatment," 153.

While tolerance is a value that can support protecting conscience, the more significant point for our purposes is that tolerance is a state interest and something that states should protect in their laws and policies. Liberal societies, at least, are committed to state neutrality about the good, which speaks in favor of tolerance.[63] Wear and colleagues claim tolerance of moral diversity is "a first principle, particularly in post-industrial, democratic societies."[64] Tolerance can promote diversity and moral reflection, which in turn promotes a healthy democracy.[65]

Tolerance is an important value and a state interest, and adding tolerance to integrity does add some weight in favor of protecting conscientious refusal generally. Nevertheless, I am unconvinced that tolerance can provide enough extra support for complicity refusal. Tolerance is valuable because it promotes diversity, moral reflection, and cooperation.[66] Yet as those who advocate tolerance recognize, it can be trumped by other values, especially when it fails to promote diversity and cooperation. Clearly, if a practice is itself intolerant (e.g., racist, sexist, etc.), it need not be respected.[67] But we need not go this far; if respecting someone's conscience "entails a substantial risk of serious illness, injury, or death to the party that disagrees with the practice, there are grounds for considering whether the practice can justifiably be tolerated."[68] I have argued above that tolerance of complicity refusal does involve these substantial risks, and it also seriously threatens crucial values of respect and autonomy. Significantly, these values are also important for a healthy democracy. Adding tolerance to the scale alongside integrity is insufficient against these competing values and does not provide the needed support for complicity refusal.

Yet perhaps we can add something further to integrity and tolerance in support of complicity refusal. Maybe those who lose their integrity when they are made complicit in activities that they believe are wrong also lose their *moral standing*—often cashed out as a moral right—to hold others accountable for similar activities. Because it is a valuable state interest to have a society in which

63 Ben-Moshe, "The Truth Behind Conscientious Objection in Medicine," 404.

64 Wear, Lagaipa, and Logue, "Toleration of Moral Diversity and the Conscientious Refusal by Physicians to Withdraw Life-Sustaining Treatment," 147.

65 Wester, "Conscientious Objection by Health Care Professionals," 430.

66 Indeed, at times, some advocates write as if tolerance is valuable *because* it protects integrity. Wear and colleagues write of objectors who are forced to be complicit feeling morally responsible for wrongdoing rather than "off the moral hook" ("Toleration of Moral Diversity and the Conscientious Refusal by Physicians to Withdraw Life-Sustaining Treatment," 150). If tolerance is valuable in part because of its role in protecting integrity, it adds even less weight when added to integrity.

67 Sulmasy, "What Is Conscience and Why Is Respect for It So Important?" 146.

68 Sulmasy, "What Is Conscience and Why Is Respect for It So Important?" 146.

individuals have the right to hold each other accountable, perhaps protecting complicity refusal is justified after all.

To understand this argument, it is useful to briefly survey the nature of moral standing and its relationship with hypocrisy and the closely related concept of complicity. Theorists writing on standing have largely seen it as a right or entitlement.[69] Even if an individual is blameworthy for wrongdoing, we cannot automatically assume that just anyone has the right to blame them. To illustrate, suppose that Nidhi disrespects her students by regularly arriving late to teach her class. She is plausibly blameworthy for wrongdoing. But if I am also regularly unapologetically late to teach my class, *I* do not have the right to blame Nidhi for her lateness. I would be hypocritical, and being hypocritical with regard to some norm or value undermines one's right to blame others for that norm or value.[70]

This is relevant because some have claimed that individuals who are required to be involved in activities to which they conscientiously object are made to be hypocritical. For instance, Gerrard writes when discussing referrals, "From this, it is easy to imagine that conscientious objectors could be viewed as judgemental hypocrites."[71] Yet even if such individuals should not rightly be called *hypocrites*, many in the literature have argued that they may plausibly be seen as *complicit*. Both hypocrisy and complicity are generally thought to undermine moral standing. Nicolas Cornell and Amy Sepinwall argue that this concern with moral standing is a state interest: "a state should care about protecting individuals' standing to engage in moral address for reasons related to the benefits of the so-called marketplace of ideas. . . . Being put in a position where one's standing to make certain claims is undermined should be viewed as an impairment of an individual's speech interests."[72] Citizens need standing to advocate for their beliefs, and there is great value in citizens being able to advocate for those beliefs freely so that society can discover the best ideas. Compromised standing impacts the equality of citizens in moral accountability, and this is something the state has an interest in protecting. The ability to engage in legitimate moral discourse is a weighty consideration, and alongside integrity and tolerance, perhaps it could provide what is needed to compete with the burdens to patients caused by complicity refusal.

69 Fritz and Miller, "A Standing Asymmetry Between Blame and Forgiveness," 766–68.

70 Cohen, "Casting the First Stone"; Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons"; Fritz and Miller, "Hypocrisy and the Standing to Blame"; and Todd, "A Unified Account of the Moral Standing to Blame."

71 Gerrard, "Is It Ethical for a General Practitioner to Claim a Conscientious Objection when Asked to Refer for Abortion?" 601.

72 Cornell and Sepinwall, "Complicity and Hypocrisy," 169.

Why do Cornell and Sepinwall think that those who are involved in some activity they find objectionable lack standing? Their argument begins with hypocrisy. Much of the literature on standing has focused on what Cornell and Sepinwall call *hypocritical moral address*, which involves some kind of blame or holding accountable for wrongdoing when one is hypocritical.[73] There are multiple possible explanations for why the hypocritical blamer lacks the right to blame, but two dominate the literature. On the moral equality view, the hypocritical blamer rejects the moral equality of persons when they are disposed to blame others without blaming themselves for relevantly similar faults. Because the right to blame is grounded in the equality of persons, however, the hypocritical blamer lacks the right to blame others for the relevant norm violation.[74] A different view, the commitment view, holds that the trouble with hypocritical blame is that such blamers are not sufficiently committed to the relevant norm they blame others for violating. If they were so committed, they would blame themselves as well as others for violating the norm. This lack of commitment to the relevant norm undermines one's right to blame others for violating the norm.[75]

Yet Cornell and Sepinwall suggest there is a weaker sort of hypocrisy than hypocritical moral address: *mere hypocritical inconsistency*. Mere hypocritical inconsistency "involves failing to conform one's conduct to one's moral judgments, but without blaming or addressing others."[76] This mere inconsistency, they suggest, does not undermine an individual's standing to blame. After all, they are not blaming anyone. But because of this inconsistency, if the merely inconsistent hypocrite *were* to blame someone, *then* they would open themselves up to the charge of hypocritical moral address. As Cornell and Sepinwall write, "mere hypocritical inconsistency is a proto version of hypocritical moral address. This suggests a reason not to be hypocritically inconsistent: it makes one's future moral address liable to being hypocritical."[77] The idea is that even

73 Cornell and Sepinwall, "Complicity and Hypocrisy," 157.

74 Fritz and Miller, "Hypocrisy and the Standing to Blame," 125–27, and "The Unique Badness of Hypocritical Blame," 546–50.

75 Riedener, "The Standing to Blame, or Why Moral Disapproval Is What It Is"; Todd, "A Unified Account of the Moral Standing to Blame"; Lippert-Rasmussen, "Why the Moral Equality Account of the Hypocrite's Lack of Standing to Blame Fails"; and Piovarchy, "Hypocritical Blame as Dishonest Signaling." Why lack of commitment to a norm undermines one's right to blame has been disputed. Todd and Riedener both see it as a fundamental fact. Piovarchy, however, suggests that blame is justified by signaling commitment to a norm, but hypocritical blame is dishonest signaling that undermines the very function of blame.

76 Cornell and Sepinwall, "Complicity and Hypocrisy," 157.

77 Cornell and Sepinwall, "Complicity and Hypocrisy," 164.

if one does not lose one's standing to blame through mere hypocritical inconsistency, such inconsistency limits one's ability to hold others accountable for wrongdoing in the future. Once one morally addresses another for wrongdoing that one has been inconsistent about oneself, one enters the realm of hypocritical moral address, and the arguments for why one's standing to blame is undermined come into play.

Importantly, complicity is also thought to undermine the standing to blame. If I am involved in some wrongdoing, I forfeit my right to blame you for the wrongdoing. After all, I am partly to blame. While it may be more objectionable to be willingly involved in wrongdoing, Cornell and Sepinwall argue that there is a weaker form of complicity: *mere involvement complicity*. This complicity is merely being involved in some wrong regardless of whether one intends to contribute to the wrong and regardless of whether one's actions could prevent the wrong from happening.[78] The connection to the wrong, however tenuous, is enough for mere involvement complicity. Just as being hypocritically inconsistent prevents one from holding others accountable on pain of engaging in hypocritical moral address, mere involvement complicity prevents one from holding others accountable on pain of becoming something akin to a hypocrite.[79] In other words, even mere involvement complicity can limit an individual's ability to morally engage in the community and blame others for certain perceived wrongdoing.

We can formalize the heart of Cornell and Sepinwall's argument and apply it in the current context of the overweighted integrity problem:

1. When the state denies complicity refusal, it forces others to be associated with behaviors that they believe are wrong.
2. If one is associated with some activity that one believes is wrong, then one's standing to hold others accountable for that behavior is undermined.
3. So when the state denies complicity refusal, it undermines the standing of its citizens to hold others accountable for behaviors that they believe are wrong.
4. The state should protect the standing of its citizens to hold others accountable for behaviors that they believe are wrong.
5. So the state should protect complicity refusal.

I have explained Cornell and Sepinwall's reasoning for the premises above. If the argument works, the state ought to keep complicity refusal in place not

78 Cornell and Sepinwall, "Complicity and Hypocrisy," 161.

79 Cornell and Sepinwall, "Complicity and Hypocrisy," 166.

merely to protect integrity but to protect the moral standing of its citizens and to ensure a flourishing society that can engage in moral discourse. In the next section, however, I will argue that the argument fails with the second premise.

## 4. THE FALL OF THE MORAL STANDING ARGUMENT

A key premise of the argument above is that being associated with some activity that one believes is wrong undermines one's standing to hold others accountable for that behavior. Cornell and Sepinwall say that one reason not to be merely hypocritically inconsistent is that "it makes one's future moral address liable to being hypocritical."[80] Something similar can be said about mere involvement complicity. If I am in some way connected to the wrong behavior—however tangentially—it might seem that I would not be entitled to criticize others for that behavior.

This raises an important question though. When is standing undermined, and why? Cornell and Sepinwall are not clear on this point, but the quote above could be read to suggest that standing is undermined *in the process of blaming*. On this picture, being weakly complicit *itself* does not undermine one's standing; that standing is lost only when one attempts to make a moral address. Alternatively, one might think that standing is undermined before ever making an address.

The first option is implausible. It suggests that one lacks the standing to do something only when one tries to do it, and not before. Yet it is difficult to see how only by engaging in *X* do I thereby lack the right to *X*. If one lacks a right to *X* only by engaging in *X*, then it is too late to lose that right. Instead, it is more plausible that the right to blame is lost *before* any address. To illustrate, an unfaithful lover might be unaware that his partner is also unfaithful. Because he is unaware, he does not blame his lover. But he lacks the standing to blame for infidelity before ever actually blaming.[81] It is not the case that the lover has the standing to blame just up to the point at which he begins to blame and then loses that right when trying to exercise it in the process of blaming.

If standing is undermined before one makes any moral address, there must be something that undermines that standing prior to the address itself. So does mere involvement complicity undermine one's standing to blame? And if so, why?

80 Cornell and Sepinwall, "Complicity and Hypocrisy," 164.

81 This is the example of Cato and Danae I have used elsewhere. See Fritz and Miller, "Two Problems of Self-Blame for Accounts of Moral Standing," 846.

If involvement complicity undermines standing, the first point to appreciate is that nearly everyone will lose standing to blame for a variety of wrongs in the world. Through politics, capitalism, trade markets, and dumb luck, each of us lies in some causal chain that could plausibly be tied back to wrongdoing.[82] To illustrate, using a variety of electronic devices makes one complicit in the mining of rare earth elements, climate change, and other environmental harms.[83] Nevertheless, most of us believe that we have the right to blame polluters, companies that contribute to climate change, or celebrities and billionaires with large carbon footprints. Residents of the United States pay taxes that may fund wars, thereby making them complicit in those wars. If this involvement complicity undermines standing, no tax-paying US resident has the right to blame their government for what they see as an unjust war.[84]

The upshot is that many of us are complicit, however weakly, in a great deal of activities that we may think are wrong. If this weak complicity undermines one's standing, many of us lack the standing to morally address others for such wrongdoing. That would be an unwelcome conclusion—especially if the state has a strong interest in ensuring citizens are entitled to morally address each other. But this conclusion on its own does not show that complicity does not undermine standing. Perhaps there is good reason to think that this unwelcome conclusion is nevertheless true. To determine that, we must turn to the most common explanations of when standing is undermined and see if they apply to the case of involvement complicity.

82 As the band Spanish Love Songs mourns in their song, "Optimism (As a Radical Life Choice)," "Can't even have my coffee without exploiting someone or making another millionaire a billionaire."

83 Balaram, "Rare Earth Elements."

84 Cornell and Sepinwall acknowledge this concern with taxes: "One might think that, by paying taxes that support a war effort one becomes complicit in that war" ("Complicity and Hypocrisy," 178n31). Nevertheless, they suggest that taxes are different from conscription:

> Because everyone pays taxes, we are all placed in a similar position, which it is hard then to view as a disability insofar as it is generally shared. . . . One might think of this difference as based on a shared understanding that none of us will treat the *de minimis*, fungible contributions of our tax dollars as undermining each other's standing, because we all know that the government will inevitably fund projects that each of us does not believe in from time to time (178n31).

Yet the idea that we are all in a similar position is precisely the point. If mere involvement complicity undermines standing, then it is not particularly useful to reply that we can just ignore everyone's undermined standing in some cases. One still lacks the right to blame in such cases. Yet this seems the wrong result: I suspect many of us would insist that we *do* have the right to blame a government for some war to which we are morally opposed, even knowing that our taxes help fund that war. The explanation for this, as I explain below, depends on an individual's attitudes.

Let us begin with the commitment view. On this view, if someone demonstrates they are not sufficiently committed to some norm, they lack the standing to blame for violations of that norm. This could apply when someone hypocritically blames others for wrongdoing but not themself. If that person were properly committed to the norm, they would blame themself for violations just as they blame others. Or it could occur when they are problematically involved in wrongdoing, since that involvement could indicate that they lack sufficient commitment to the norm.

Does involvement with some norm violation show that one lacks sufficient commitment to the norm? Not necessarily. Patrick Todd, a proponent of the commitment view, writes, "It is, at most, only a particular *kind* of involvement that removes standing. . . . Involvement removes standing only when it indicates a lack of commitment to the values that would condemn the wrongdoer's actions."[85] The commitment at issue requires "*endorsement* of the value as a genuine value" and "at least some degree of *motivation* to act in accordance with the value."[86] The better question to ask, then, is whether one can be committed to some norm (or value) while still being complicit in some violation of that norm.

The answer is plausibly affirmative—especially if one is *unwillingly* complicit in the violation. Those who are legally compelled to be complicit in a norm violation can nevertheless strongly endorse that norm and remain motivated to act in accordance with it. This is precisely the situation many health care professionals might find themselves in if we no longer protect complicity refusal. It is not as if someone who thinks that performing an abortion violates a norm against killing will no longer endorse such a norm simply by cleaning instruments used in the procedure. Even those who provide information on abortion can endorse norms that forbid it. Similarly, they will retain their motivation to act in accordance with the norm. The very fact that their involvement is unwilling indicates their commitment to the norm.

At this point, one might object that sufficient commitment is quite strong: perhaps "one must be unassailably free of taint from a wrong if one is to condemn others for it."[87] Phrased differently, one might think that an unblemished moral record with respect to some norm is required to show that one is committed to a norm, yet involvement complicity taints that moral record.

It is important to stress that one can remain committed to a norm without being a moral saint who perfectly complies with the norm. Consider the case of

85 Todd, "A Unified Account of the Moral Standing to Blame," 355. Todd frames his view in terms of commitment to *values* rather than commitment to *norms*, but nothing of substance hangs on this distinction.

86 Todd, "A Unified Account of the Moral Standing to Blame," 355.

87 Cornell and Sepinwall, "Complicity and Hypocrisy," 168.

Nina, who claims to be deeply committed to a norm against losing one's temper when a child misbehaves. Nina takes various steps to ensure she abides by this norm, including meditation and exercise. Yet one day, she loses her temper with her child, in part due to factors out of her control. Even when she fails to comply with the norm on this one occasion, it would be implausible to claim that Nina is not seriously committed to the norm against losing one's temper.[88] And in this case, Nina directly violates the norm herself; she is not merely associated with the violation. While we may disagree about exactly how compliant with a norm one must be in order to be sufficiently committed to it, moral perfection is, in my view, clearly too high of a bar.

If individuals need not be morally perfect to be committed to a norm, then if there is any moral taint that comes with merely being involved in some potential wrongdoing, it might not be problematic for the commitment view. Being willingly complicit in something one claims to believe is wrong may suggest that one lacks the relevant commitment. But being unwillingly complicit does not suggest that one does not endorse the relevant value or lacks the motivation to uphold it. This shows the importance of attitudes and beliefs in determining an individual's standing; it is not merely a matter of whether an individual is somehow connected to some wrongdoing. The citizen who pays her taxes knowing that those taxes fund a war to which she is morally opposed may still endorse pacifist values and norms. Perhaps she shows this endorsement by participating in protests, calling her representatives, and actively writing about why the war is wrong. She pays taxes only because they are compulsory—but what demonstrates one's commitment to a norm is what one does *freely*. Compulsory actions reveal little about the norms one endorses internally.

In sum, if standing requires commitment to a norm, complicity does not necessarily undermine that standing. One can remain sufficiently committed while being complicit, depending on one's attitudes and what external forces are at play. There is simply no good motivation for understanding the necessary commitment to a norm as so impossibly high that it means one cannot be in any way involved or associated with anything that violates that norm.

If involvement complicity does not necessarily show that one is not sufficiently committed to a norm and thereby lacks the standing to blame for that norm, then premise two cannot be supported with the commitment view. Yet the moral equality view stands as the other chief explanation for undermined standing. If mere involvement in wrongdoing shows that one rejects the moral equality of persons, then premise two could be supported in that way.

88 Fritz and Miller, "Two Problems of Self-Blame for Accounts of Moral Standing," 840.

Currently, the most developed version of the moral equality view is the one I have advocated with Daniel Miller.[89] We argue that an individual's standing is undermined when they are unfairly disposed to blame differentially for violations of some norm. The reason for this is that such unfair differential blaming dispositions implicitly reject the moral equality of persons, yet that equality of persons is what grounds the right to blame in the first place.

It is important to highlight the role of dispositions in explaining why standing is undermined on this view. Miller and I hold that merely engaging in behavior you have condemned as wrong does not thereby undermine your standing.[90] What matters are your attitudes regarding that behavior, both towards yourself and others. Inconsistency in one's professed values and behaviors does not automatically show that one has implicitly rejected the moral equality of persons. For example, if an akratic vegetarian eats a burger but feels guilty and blames themself for wrongdoing, they do not reject the equality of persons. They treat themself just the same as they would others who eat meat.

As discussed above, if we reject complicity refusal, health care professionals would be *compelled* to be involved with actions that they believe are wrong and thereby would be compelled to be inconsistent in their actions and attitudes. Just as in the case of the commitment view, this compulsion is significant because it likely indicates a lack of any problematic differential blaming dispositions.

Consider three different types of agents. Julien holds very high standards and thinks that being in any way involved with perceived wrongdoing is blameworthy. Although Julien is compelled by the state to be complicit in something she believes is wrong, she is disposed to blame herself just as she is disposed to blame others. She feels her integrity is tarnished due to her complicity, and this produces guilt and self-blame. Julien is consistent in her blaming dispositions and so maintains the standing to blame despite being complicit, because she holds herself to the same standards as others.

A second agent, Lucy, is less strict than Julien. Lucy believes that compulsion by the state is a reasonable excuse for engaging in mere tangential involvement in some perceived wrongdoing. Accordingly, when Lucy is compelled by the state to be complicit in something she believes is wrong, she is not disposed to blame herself. Yet Lucy also lacks any unfair differential blaming disposition, because she is not disposed to blame anyone *else* who is compelled to be complicit either. While Lucy may be disposed to blame those who *willingly*

89 Fritz and Miller, "Hypocrisy and the Standing to Blame" and "The Unique Badness of Hypocritical Blame."

90 Fritz and Miller, "Hypocrisy and the Standing to Blame," 121.

violate the norm, she grants the same grace to others that she grants to herself regarding compulsion. Consequently, she too retains the standing to blame on a moral equality account, because she lacks any differential blaming disposition.

The final agent is Phoebe. In contrast to Julien and Lucy, Phoebe is disposed to let herself off the hook for being compelled to be complicit in something she believes is wrong but nevertheless is disposed to blame others who are similarly compelled to be complicit in wrongdoing. Phoebe *does* lack the standing to blame others, as she has an unfair differential blaming disposition.

While there could be health care professionals like Phoebe, such agents seem straightforwardly hypocritical, and we should expect them to lack the standing to hold others accountable for the relevant norm violation. But more likely are agents like Julien or Lucy, who are consistent in their blaming dispositions. Some individuals will see compulsion as a reasonable excuse for everyone and will let themselves and others off the hook as a result. Others will refuse to see compulsion as an excuse and so will blame anyone who is compelled to be complicit in wrongdoing—themselves included. After all, the sorts of agents who see mere involvement in some activity as a threat to their integrity will also probably blame themselves for being involved in that activity. Either way, such individuals do not lack the moral standing to blame on the moral equality account, and premise two remains unsupported.[91]

Cornell and Sepinwall are quick to reject the notion that compulsion can protect one's standing, especially when we consider the excuses offered by individuals during trials after the fall of the Nazi regime or apartheid.[92] But even if one maintains the standing to blame others, this does not imply that they are not guilty of any wrongdoing. We can condemn the actions of those who were complicit in Nazi Germany or apartheid without holding that their standing to blame others is undermined. What matters for standing is not simply compulsion but the consistent blaming dispositions of the agent; if someone blames themself just as they blame others for their unwilling involvement in the Nazi regime, they maintain their standing to blame. Nevertheless, it is plausible that the unwillingly complicit should have done more to resist these great atrocities. We need not deny their standing to condemn their complicity.[93]

91 Notably, an agent's standing is not completely at the mercy of the state on the moral equality account. Individuals who have lost their standing via differential blaming dispositions could regain that standing simply by coming to be disposed to blame themselves the same as others who are similarly compelled.

92 Cornell and Sepinwall, "Complicity and Hypocrisy," 168.

93 It is worth emphasizing that there are relevant differences between complicity in a genocidal regime and complicity in providing medically accepted but morally contested care. The

Cornell and Sepinwall do consider that these attitudes could be significant: "The compelled actor might well feel terrible, acknowledge his fault, and blame himself for it. Shouldn't this person's standing then remain intact? Perhaps. But even if this person treats others just as he treats himself, no one else will know this."[94] This is an odd response, however. Whether one *believes* someone has blamed themself clearly has no bearing on whether they have *actually* blamed themself. Someone may have standing to blame regardless of whether anyone knows this. If the state is to have an interest in protecting moral standing, it must be *actual* moral standing, not simply what people might *believe* about moral standing. It would be portentous to make accommodations at great cost to society merely because some individuals *think* their standing is undermined when it in fact is not.

In sum, we were hunting for something valuable to place on the scale alongside integrity and tolerance that might justify protecting complicity refusals. Moral standing was a promising candidate. But one must first show that being associated with some activity that one believes is wrong, even if unwilling, undermines one's standing to hold others accountable for that behavior. There is no good reason to believe this. Neither of the leading explanations of undermined standing, the commitment view and the moral equality view, support it. Without the support of that crucial premise, there is no reason to believe that moral standing is actually undermined when the state compels individuals to be associated with behavior that they believe is wrong.

## 5. CONCLUSION

Current policies protecting complicity refusal are unjustified. Significant values that are integral to biomedical ethics, such as autonomy, beneficence, and nonmaleficence, are on the line, as well as patient rights and negative consequences for society. Integrity alone cannot compete with these values, and tolerance adds little additional weight. If the moral standing of health care professionals would be widely undermined without complicity refusal, this may be enough to tip the scales, since the state has an interest in ensuring citizens can hold each other to account. Yet this appeal to moral standing fails. Complicity refusal is, in the end, unjustified.

What are the policy implications for such a conclusion? First, conscience clauses need to be rewritten to exclude such broad complicity refusals. How

---

stakes are much higher in the former case, so there may not be the same duty to resist in the latter set of cases.

94 Cornell and Sepinwall, "Complicity and Hypocrisy," 168–69.

restrictive these clauses should be depends on empirical evidence regarding how often patients are denied relevant information about their care and the negative consequences of this ignorance. It may be reasonable to accommodate direct conscientious refusal while still respecting patient autonomy and well-being, provided patients can nevertheless receive the information and care they need in a timely manner. Whether this is possible for some subset of complicity refusals is unclear. Health care professionals, philosophers, and policymakers should all be involved in that discussion. One thing is clear though: the way many policies are currently worded is too broad.

Second, institutions should provide a method of making clear that some individual is only involved in some activity because of a legal requirement. Even Cornell and Sepinwall make this suggestion: "where the state imposes a contested legal requirement on someone who objects, it might incur an affirmative duty to make clear to others that the objector complies only because she is legally compelled to do so."[95] This allows those who object to stand apart from those who willingly participate, ensuring the objectors do not unfairly condemn other compelled actors and thereby undermine their own standing. It also provides a way for objectors to credibly demonstrate their moral beliefs in some manner, even if they must act in ways that they see as inconsistent with those beliefs. It may even minimize feelings of guilt or remorse or mitigate the threat to an individual's moral identity. There are various methods that institutions could adopt to share such information, including maintaining a database or encouraging objectors to wear some token to signal their objection.

The current state of conscience policy in US health care is unjust. In many states, policies protect the integrity of health care professionals to such an extent that professionals need not be involved in any way with care to which they object. Yet these policies leave patients without sufficient information about their own care, and the policies can result in serious negative health outcomes for patients like Tamesha Means. Despite the value of integrity, it cannot compete against autonomy, benevolence, and nonmaleficence. Valuing patients requires restricting complicity refusal.[96]

*University of Mississippi*
*kgfritz@olemiss.edu*

95 Cornell and Sepinwall, "Complicity and Hypocrisy," 171.

96 Thanks to Sara Kolmes, Dan Miller, and an audience at the Western Michigan University Medical Humanities Conference for valuable discussion that helped to shape my ideas into this paper. I am also grateful for the feedback from two anonymous referees that allowed me to sharpen and strengthen the arguments in the paper.

## REFERENCES

Balaram, V. "Rare Earth Elements: A Review of Applications, Occurrence, Exploration, Analysis, Recycling, and Environmental Impact." *Geoscience Frontiers* 10, no. 4 (2019): 1285–303.

Bayles, Michael D. "A Problem of Clean Hands: Refusal to Provide Professional Services." *Social Theory and Practice* 5, no. 2 (1979): 165–81.

Beauchamp, Tom L., and James F. Childress. *Principles of Biomedical Ethics*. 8th ed. Oxford University Press, 2019.

Ben-Moshe, Nir. "Conscientious Objection in Medicine: Making it Public." *HEC Forum* 33, no. 3 (2021): 269–89.

———. "Internal and External Paternalism." *Canadian Journal of Philosophy* 52, no. 6 (2022): 673–87.

———. "The Truth Behind Conscientious Objection in Medicine." *Journal of Medical Ethics* 45, no. 6 (2019): 404–10.

Benn, Piers. "Conscience and Health Care Ethics." In *Principles of Health Care Ethics*, edited by Richard E. Ashcroft, Angus Dawson, Heather Draper, and John McMillan. Wiley, 2007.

Birchley, Giles. "A Clear Case for Conscience in Healthcare Practice." *Journal of Medical Ethics* 38, no. 1 (2012): 13–17.

Blustein, Jeffrey. "Doing What the Patient Orders: Maintaining Integrity in the Doctor-Patient Relationship." *Bioethics* 7, no. 4 (1993): 289–314.

Brock, Dan. "Conscientious Refusal by Physicians and Pharmacists: Who Is Obliged to Do What, and Why?" *Theoretical Medicine and Bioethics* 29, no. 3 (2008): 187–200.

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

Card, Robert F. "Conscientious Objection and Emergency Contraception." *American Journal of Bioethics* 7, no. 6 (2007): 8–14.

Chang, Ruth. "Value Incomparability and Incommensurability." In *The Oxford Handbook of Value Theory*, edited by Iwao Hirose and Jonas Olson. Oxford University Press, 2015.

Clarke, Steve. "Conscientious Objection in Healthcare, Referral and the Military Analogy." *Journal of Medical Ethics* 43, no. 4 (2017): 218–21.

Cohen, G. A. "Casting the First Stone: Who Can, and Who Can't, Condemn the Terrorists? *Royal Institute of Philosophy Supplement* 58, no. 1 (2006): 113–36.

Combs, Michael P., Ryan M. Antiel, Jon C. Tilburt, Paul S. Mueller, and Farr A. Curlin. "Conscientious Refusals to Refer: Findings from a National

Physician Survey." *Journal of Medical Ethics* 37, no. 7 (2011): 397–401.
Cornell, Nicolas, and Amy Sepinwall. "Complicity and Hypocrisy." *Politics, Philosophy and Economics* 19, no. 2 (2020): 154–81.
Curlin, Farr A., Ryan E. Lawrence, Marshall H. Chin, and John D. Lantos. "Religion, Conscience, and Controversial Clinical Practices." *New England Journal of Medicine* 356, no. 6 (2007): 593–600.
Davis, Shoni, Vivian Schrader, and Marcia J. Belcheir. "Influencers of Ethical Beliefs and the Impact on Moral Distress and Conscientious Objection." *Nursing Ethics* 19, no. 6 (2012): 738–49.
Dresser, Rebecca. "Professionals, Conformity, and Conscience." *Hastings Center Report* 35, no. 6 (2005): 9–10.
Dworkin, Ronald. *Life's Dominion*. Harvard University Press, 1993.
Fritz, Kyle G., and Daniel Miller. "Hypocrisy and the Standing to Blame." *Pacific Philosophical Quarterly* 99, no. 1 (2018): 118–39.
———. "A Standing Asymmetry Between Blame and Forgiveness." *Ethics* 132, no. 4 (2022): 759–86.
———. "Two Problems of Self-Blame for Accounts of Moral Standing." *Ergo* 8, no. 54 (2022): 833–56.
———. "The Unique Badness of Hypocritical Blame." *Ergo* 6, no. 19 (2019): 545–69.
Gerrard, J. W. "Is It Ethical for a General Practitioner to Claim a Conscientious Objection When Asked to Refer for Abortion?" *Journal of Medical Ethics* 35, no. 10 (2009): 599–602.
Hill, Daniel J. "Abortion and Conscientious Objection." *Journal of Evaluation in Clinical Practice* 16, no. 2 (2010): 344–50.
Kaye, Julia, Brigitte Amiri, Louise Melling, and Jennifer Dalven. *Health Care Denied: Patients and Physicians Speak Out About Catholic Hospitals and the Threat to Women's Health and Lives*. American Civil Liberties Union, 2016. https://www.aclu.org/sites/default/files/field_document/healthcaredenied.pdf.
Lawrence, Ryan E., and Farr A. Curlin. "Autonomy, Religion and Clinical Decisions: Findings from a National Physician Survey." *Journal of Medical Ethics* 35, no. 4 (2009): 214–18.
Lippert-Rasmussen, Kasper. "Why the Moral Equality Account of the Hypocrite's Lack of Standing to Blame Fails." *Analysis* 80, no. 4 (2020): 666–74.
Magelssen, Morten. "When Should Conscientious Objection Be Accepted?" *Journal of Medical Ethics* 38, no. 1 (2012): 18–21.
May, Thomas, and Mark P. Aulisio. "Personal Morality and Professional Obligations: Rights of Conscience and Informed Consent." *Perspectives in Biology and Medicine* 52, no. 1 (2009): 30–38.

McLeod, Carolyn. *Conscience in Reproductive Health Care: Prioritizing Patient Interests*. Oxford University Press, 2020.

Minerva, Francesca. "Conscientious Objection, Complicity in Wrongdoing, and a Not-So-Moderate Approach." *Cambridge Quarterly of Healthcare Ethics* 26, no. 1 (2017): 109–19.

National Women's Law Center. "Below the Radar: Health Care Providers' Religious Refusals Can Endanger Pregnant Women's Lives and Health." 2011. https://nwlc.org/wp-content/uploads/2015/08/nwlcbelowtheradar2011.pdf.

Piovarchy, Adam. "Hypocritical Blame as Dishonest Signaling." *Australasian Journal of Philosophy* (forthcoming). Available at https://philpapers.org/rec/PIOHBA.

Pope, Thaddeus Mason. "Conscience Clauses and Conscientious Refusal." *Journal of Clinical Ethics* 21, no. 2 (2010): 163–80.

Ramondetta, Lois, Alaina Brown, Gwyn Richardson, et al. "Religious and Spiritual Beliefs of Gynecologic Oncologists May Influence Medical Decision Making." *International Journal of Gynecological Cancer* 21, no. 3 (2011): 573–81.

Riedener, Stefan. "The Standing to Blame, or Why Moral Disapproval Is What It Is." *Dialectica* 73, nos. 1–2 (2019): 183–210.

Robinson, Michael. "Voluntarily Chosen Roles and Conscientious Objection in Health Care." *Journal of Medical Ethics* 48, no. 10 (2022): 718–22.

Sawicki, Nadia. "The Conscience Defense to Malpractice." *California Law Review* 108 (2020): 1255–316.

———. "Mandating Disclosure of Conscience-Based Limitations on Medical Practice." *American Journal of Law & Medicine* 42, no. 1 (2016): 85–128.

Schmidtz, David. "Value in Nature." In *The Oxford Handbook of Value Theory*, edited by Iwao Hirose and Jonas Olson. Oxford University Press, 2015.

Schuklenk, Udo. "Conscientious Objection in Medicine: Private Ideological Convictions Must Not Supercede Public Service Obligations." *Bioethics* 29, no. 5 (2015): ii–iii.

Sepinwall, Amy J. "Conscientious Objection, Complicity, and Accommodation." In *Law, Religion, and Health in the United States*, edited by Holly Fernandez Lynch, I. Glenn Cohen, and Elizabeth Sepper. Cambridge University Press, 2017.

Shahvisi, Arianne. "Conscientious Objection: A Morally Insupportable Misuse of Authority." *Clinical Ethics* 13, no. 2 (2018): 82–87.

Stahl, Ronit, and Ezekiel Emanuel. "Physicians, Not Conscripts: Conscientious Objection in Health Care." *New England Journal of Medicine* 376, no. 14 (2017): 1380–85.

Sulmasy, Daniel P. "What Is Conscience and Why Is Respect for It So Important?" *Theoretical Medicine and Bioethics* 29, no. 3 (2008): 135–49.

Todd, Patrick. "A Unified Account of the Moral Standing to Blame." *Noûs* 53, no. 2 (2019): 347–74.

Uttley, Louis, Sheila Reynertson, Lorraine Kenny, and Louise Melling. "Miscarriage of Medicine: The Growth of Catholic Hospitals and the Threat to Reproductive Health Care." American Civil Liberties Union and MergerWatch Project, 2013. https://www.aclu.org/sites/default/files/field_document/growth-of-catholic-hospitals-2013.pdf.

Wallace, R. Jay. "Hypocrisy, Moral Address, and the Equal Standing of Persons." *Philosophy and Public Affairs* 38, no. 4 (2010): 307–41.

Wear, Stephen, Susan Lagaipa, and Gerald Logue. "Toleration of Moral Diversity and the Conscientious Refusal by Physicians to Withdraw Life-Sustaining Treatment." *Journal of Medicine and Philosophy* 19, no. 2 (1994): 147–59.

Wester, Gry. "Conscientious Objection by Health Care Professionals." *Philosophy Compass* 10, no. 7 (2015): 427–37.

Wicclair, Mark. "Commentary: Special Issue on Conscientious Objection." *HEC Forum* 33, no. 3 (2021): 307–24.

———. *Conscientious Objection in Health Care: An Ethical Analysis*. Cambridge University Press, 2011.

———. "Conscientious Objection in Healthcare and Moral Integrity." *Cambridge Quarterly of Healthcare Ethics* 26, no. 1 (2017): 7–17.

———. *Conscientious Objection in Medicine*. Cambridge University Press, 2024.

*Journal of Ethics and Social Philosophy*
VOL. 29, NO. 2 · JANUARY 2025
*https://doi.org/10.26556/jesp.v29i2.3376*


# CRIME, PUBLIC HEALTH, AND INHUMANE OBJECTIVITY

*Nadine Elzein*

THE IDEA that we should reject retributivism and treat crime as a public health problem seems at least *prima facie* like an appealing stance. In recent years, it has been forcefully defended by Gregg Caruso and Derk Pereboom.[1] Many of the basic principles of this view have also been defended by Erin Kelly.[2]

Plausibly, retributive policies presuppose *basic desert*, and there are reasons to doubt that agents can be assumed to basically deserve punishment for their wrongdoings. But even those who are *not* doubtful about basic desert may think that public protection is a more important goal than enacting retribution, and these goals often conflict; highly retributive systems typically do a worse job with respect to public protection than more forward-focused systems.[3]

Caruso puts forward various policy recommendations inspired by the public health model.[4] These are supported by both evidence and common sense. There are, unsurprisingly, few critiques of these recommendations in the literature. Yet whenever discussion arises about the possibility of crime being treated as a public health problem, I find that the idea meets with resistance. For many, this suggestion rings alarm bells. This is partly because the project of

1 Caruso, "Free Will Skepticism and Criminal Behavior," *Public Health and Safety*, and *Rejecting Retributivism*; Caruso and Pereboom, "A Non-punitive Alternative to Retributive Punishment"; Pereboom and Caruso, "Hard-Incompatibilist Existentialism"; and Pereboom, *Living Without Free Will*, "Incapacitation, Reintegration, and Limited General Deterrence," and *Free Will, Agency, and Meaning in Life*.

2 Kelly, *The Limits of Blame* and "Criminal Justice Without Retribution."

3 I am more optimistic about the possibility of basic desert than Caruso or Pereboom. I think it is possible that we are sometimes retributively responsible for our choices and on those occasions may be blamable to the extent of the moral disparity between the alternatives we could have chosen between. See Elzein, "Undetermined Choices, Luck and the Enhancement Problem." But both the certainty and the extent of freedom are, on this model, more limited than ordinarily assumed. Following Vargas, I am skeptical about whether we could "trace" responsibility for all of our choices back to occasional choices for which we are directly responsibility. See Vargas, "The Trouble with Tracing."

4 Caruso, "Free Will Skepticism and Criminal Behavior" and *Public Health and Safety*.

treating crime as a health problem has a dark history and partly because of the worry that there is something impersonal or dehumanizing about approaches to crime that weaken the emphasis on personal responsibility. Nonetheless, I will argue that this fear is misplaced.

In what follows, I am going to argue for five claims:

1. There is a difference between taken responsibility and retributive desert. Skepticism about retributive desert does not entail skepticism either about the existence of taken responsibility or about its moral importance.
2. The ability to take responsibility is essential to our sense of personhood, and when we undermine the abilities that underscore taken responsibility or prevent an agent from having the opportunity to take responsibility, this is commonly experienced as dehumanizing, so practices that do either unnecessarily are unethical.
3. Skepticism about retributive desert (as entailed by the public health model and Caruso's policy suggestions) is not dehumanizing in any comparable way, except where it is (needlessly) coupled with skepticism about the existence or moral importance of taken responsibility.
4. Medical approaches to crime have often been unethical historically because medical practices *in general* have often been unethical. They have often involved unnecessarily undermining an agent's capacity or opportunity to exercise taken responsibility.
5. Instead of rejecting a public health approach to crime, we should seek to take a more ethical approach to public health—one that reflects a respect for taken responsibility and therefore avoids practices that are dehumanizing (both in the context of crime and in that of public health more broadly).

In section 1, I will give a brief outline of the Public Health Quarantine Model and Caruso's policy suggestions. In section 2, I will argue for claims 1 and 2: I will show that impersonal and dehumanizing treatment comes from undermining taken responsibility, and this need not follow from skepticism about retributivism. In section 3, I will address some objections to nonretributive and public health models, all of which broadly draw on accusations that such treatment would be in some way dehumanizing, and I will argue for claim 3: skepticism about retributive desert, Caruso's policy suggestions in particular, need not be dehumanizing in the way skeptical approaches to responsibility are often accused of being. In section 4, I will briefly consider the history of the association between crime and public health, making a case for 4, the claim that unethical practices in medicine have been common but are not uniquely

associated with medicalizing deviance or criminality. In section 5, I will argue for claim 5: we ought to approach medicine more ethically instead of excluding criminality from the realm of public health.

## 1. THE PUBLIC HEALTH QUARANTINE MODEL

For Caruso and Pereboom, the public health model of dealing with crime is motivated by free will skepticism.[5] What both take to be centrally in dispute between free will skeptics and their opponents is *basic desert*. This is the sort of responsibility that is relevant to desert of retributive blame and praise. There are various reasons why we might blame or praise an agent. But insofar as they are *basically deserving*, our reasons do not rest on any further good that might result from it. It is skepticism about responsibility in *this* sense that motivates the Public Health Quarantine Model and Caruso's policy suggestions.

In his 2017 book *Public Health and Safety*, Caruso gives a thorough analysis of the social determinants of crime and public health, drawing on considerable empirical evidence.[6] He proposes eight areas in which we could adopt policies that would enable us to deal with crime more effectively without relying on the assumption of basic retributive desert. These are summarized at the end of his discussion as follows:

1. "Invest in programs and policies aimed at reducing poverty, homelessness, abuse, and domestic violence."
2. "Increase funding for mental health services with a focus on the early and active treatment of mental illness."
3. "Secure universal access to affordable and consistent healthcare for all."
4. "Reject retributivism and purely punitive approaches to criminal justice and shift the focus to *prevention*, *rehabilitation*, and *reintegration*."
5. "End all policies that disenfranchise ex-offenders, making it more difficult for them to reintegrate back into society."
6. "Prioritize and properly fund education, especially in low-income areas, and support educational programs in prison."

5 Caruso, "Free Will Skepticism and Criminal Behavior," *Public Health and Safety*, and *Rejecting Retributivism*; Caruso and Pereboom, "A Non-Punitive Alternative to Retributive Punishment"; Pereboom and Caruso, "Hard-Incompatibilist Existentialism"; and Pereboom, *Living Without Free Will*, "Incapacitation, Reintegration, and Limited General Deterrence," and *Free Will, Agency, and Meaning in Life*.

6 Caruso, *Public Health and Safety*.

7. "Adopt policies that protect the environmental health of our communities by combating climate change, protecting air and water, and reducing/eliminating harmful toxins."
8. "Research more effective interventions and rehabilitation strategies for psychopathy."[7]

Since present public health failures often worse affect those already unfairly disadvantaged by factors such as class, race, ethnicity, lifestyle, and culture, Caruso proposes policies that are informed by an ethical awareness of issues of social justice. His ethical framework involves taking a capabilities approach to public well-being, grounded in Amartya Sen's idea of enabling people to function so as to protect the substantive freedom a person has "to lead the kind of life he or she has reason to value."[8] Acknowledging this ethical commitment will be important for what follows.

Discomfort about treating crime as a public health problem apparently does not derive from unease about the specific reforms Caruso recommends, which have attracted little criticism. Some critics proclaim to be broadly supportive of some of these policy reforms but skeptical about whether Caruso's ethical commitments are consistent with his skepticism about free will.[9] And while few have attacked Caruso's view directly, there is a body of criticism predating Caruso's work that continues to be influential within free will literature and that captures the basic motivations for continued unease about the association between crime and public health. These will be discussed in section 4.

When theorists talk about moral responsibility, it is not always clear what the term is taken to mean. While Pereboom and Caruso are careful to specify that they are concerned solely about basic retributive desert, commentators do not always clearly address the relation between retributive responsibility and broader uses of the term 'responsible'. In the following section, I will distinguish three different senses in which we might use the term 'responsibility' and will make a case for supposing that one variety of responsibility—what I call *taken responsibility*, or future-directed commitment—need not stand or fall with basic retributive desert.

7 Caruso, *Public Health and Safety*, 20, 21, 24, 26.

8 Sen, *Development as Freedom*, 87, quoted in Caruso, *Public Health and Safety*, 19.

9 Levy, "Let's Not Do Responsibility Skepticism"; and Lemos, *Free Will's Value*, 148–72.

## 2. RETRIBUTIVE DESERT VERSUS FUTURE-DIRECTED COMMITMENT

### *2.1. Three Types of Responsibility*

The terms 'free will' and 'moral responsibility' lend themselves to several interpretations. Watson contrasts two notions of responsibility. The "self-disclosure view" captures *attributability* or the *aretaic* face of responsibility. In contrast, *accountability* captures the sort of responsibility that justifies desert of praise or blame.[10] We may contrast both of these with a *virtue* sense of the term, frequently associated with "taking responsibility" and less frequently discussed in relation to free will.

#### *2.1.1. Accountability or Retributive Desert*

Being responsible in the accountability sense entails being basically deserving of blame or praise. Holding someone *retributively responsible* entails blaming and praising or punishing and rewarding *just* on the basis that it is deserved. This is what would be required to justify a retributive stance: we are justified in punishing wrongdoers *just* on the basis that they deserve it. Caruso and Pereboom endorse skepticism solely about responsibility in this sense. There is considerable disagreement among philosophers about what conditions an agent must meet in order to have this variety of responsibility.[11]

#### *2.1.2. Attributability or Self-Disclosure*

The features that ground attributability are reflected in various compatibilist (or partially compatibilist) accounts. Actions may be attributable to agents to varying degrees, depending on such features as whether

- the agent performed the action deliberately,[12]
- the agent was acting on desires that she endorsed through second-order volitions,[13]
- the agent's second-order volitions reflected her deepest or most wholeheartedly embraced system of values,[14]

10 Watson, "Two Faces of Responsibility."

11 Conditions for retributive responsibility range from supposing it merely requires that our actions are conscious, intentional, rational, and uncompelled (Morse, "Compatibilist Criminal Law") to supposing that it requires us to be "miracle-working godlike beings" (Waller, "Virtue Unrewarded," 433–34).

12 Hobbes, "Of Liberty and Necessity"; and Hume, *A Treatise of Human Nature*, 2.3.1–2, 257–65 and *An Enquiry Concerning Human Understanding*, sec. 8, 148–64.

13 Frankfurt, "Freedom of the Will and the Concept of a Person."

14 Watson, "Free Agency."

- the agent's deeper values were not misguided or were at least shaped by reasoned reflection,[15] or
- the agent's mechanism of decision-making was adequately reasons-responsive.[16]

These features capture whether an agent's actions express her true intentions, character, and values, and whether these values are embraced through rational reflection as opposed to being picked up thoughtlessly or through blind indoctrination. This indicates that an agent's choices are a true reflection of the sort of person she is.

Classically, agents are exempted or excused from responsibility in the attributability sense either because the *agent* lacks the general capacities required to perform actions that are attributable to her (e.g., she lacks the ability to reason about her values or to reliably translate her values into choices and actions) or because the *action* does not reflect her true character and values. The former category may include some addicts, the severely mentally ill, or children. The latter may include actions performed accidentally, involuntarily, or through ignorance.

It is less clear that there is a distinctive attributability sense in which we might hold agents responsible, though attributability seems essential to certain practices. For example, rewards and punishments aimed solely at incentivizing good behavior or disincentivizing bad behavior make sense only when aimed at agents who meet conditions of attributability. We usually cannot incentivize someone to do something involuntary.

### *2.1.3. Taken Responsibility or Future-Directed Commitment*

Gaden contrasts the virtue sense of responsibility with the *capacity* sense.[17] Watson's two senses of responsibility both seem to fall into the capacity category. The capacity senses of 'responsible' are contrasted with 'not responsible', whereas the virtue sense of 'responsible' is contrasted with 'irresponsible'.[18] If we think about being able to take responsibility on a model akin to developing a virtue, this raises questions about how we develop this virtue, how we educate children to develop it, and how those who have developed a corresponding vice might cultivate it.

Taking responsibility is normally done prospectively and hence is predominantly forward-looking in a way that retributive responsibility is not. Doret de Ruyter notes that "a person who takes responsibility for the well-being of

15 Wolf, *Freedom Within Reason*, especially 67–93.

16 Fischer and Ravizza, *Responsibility and Control.*

17 Gaden, "Rehabilitating Responsibility."

18 Gaden, "Rehabilitating Responsibility," 27.

another tries to establish something, whereas the person who is responsible for her action is accountable for something she has already done or for something she should have done."[19] Bruce Waller uses the term 'take-charge responsibility' for something like this virtue sense.[20] Pereboom and Caruso also explicitly distinguish taking responsibility in the sense of sincerely committing to a task with the sort of responsibility relevant to basic desert of praise and blame.[21]

An agent "takes responsibility" when they exhibit *future-directed commitment.* An agent exhibits future-directed commitment only insofar as they are willing and able to commit prospectively, sincerely, and conscientiously to a project or aim. When we attribute the virtue of being a responsible person to someone, we are saying that that person reliably exhibits future-directed commitment, particularly where they are morally required to. When we describe someone as an irresponsible person, we are saying that they do not reliably exhibit future-directed commitment, especially where this involves moral negligence. A responsible person is one who can be relied upon to take responsibility when it is called for.

I will use the terms 'future-directed commitment' and 'taken responsibility' interchangeably. (The latter is more in keeping with common usage, while the former better marks the distinction between this concept and responsibility of the sort usually in question in disputes about free will.)

De Ruyter outlines a number of abilities required for an agent to count as responsible in the virtue sense. These include *rationality*, "because one has to be able to interpret the needs of others and reflect on one's possible responses"; *caring* about the needs of others; and having the *willpower* to act on this, even when we have countervailing interests.[22] When we talk about holding an agent responsible in the sense that corresponds to this sort of responsibility, this involves expecting the agent to take responsibility, e.g., expecting her to exhibit a future-directed commitment to behave better in future or to make amends for something done previously. This expectation need not involve retributive blame. If we call it "blame" at all, it may be something closer to T. M. Scanlon's *nonpunitive* form of blame.[23] But it is better captured by Hannah Pickard's notion of *responsibility without blame*. Pickard argues that this way of holding agents responsible is effective in improving behavior in both therapeutic

19 De Ruyter, "The Virtue of Taking Responsibility," 26.

20 Waller, *Against Moral Responsibility*, 105.

21 Pereboom, *Living Without Free Will*, xxi; and Caruso and Pereboom, *Moral Responsibility Reconsidered*, 3–4.

22 De Ruyter, "The Virtue of Taking Responsibility," 28–30.

23 Scanlon, "Interpreting Blame."

contexts and criminal justice contexts.[24] When we hold someone responsible in this sense, the goal is not to blame them but to foster the sorts of reflection that might enable them to better exhibit future-directed commitment.

Two questions arise here. The first is the question of what the relation is between future-directed commitment and the two capacity senses of responsibility. The second is the question of whether we can endorse skepticism about retributive desert without endorsing skepticism about one or both of the others.

### *2.2. The Relation Between Senses of 'Responsibility'*

I want to suggest that while some degree of attributability is necessary in order for an agent to be able to take responsibility, these two sorts of responsibility are only weakly connected. And it is not necessary at all that an agent meets the conditions of accountability or basic retributive desert in order to exhibit future-directed commitment of the sort required for taken responsibility.

Waller points out that while it is often assumed that take-charge responsibility suffices for being responsible in the sense that justifies blame and praise, this assumption is unjustified. Establishing that someone has take-charge responsibility still leaves open the question of whether they would be blameworthy or praiseworthy for what they have done.[25] It might seem, on the face of it, that one cannot be entailed by the other since one of these essentially involves a future-directed mindset while the other is backward-looking. While the future-directed commitment that characterizes taking responsibility is something we exercise prospectively, it can also have a backward-looking aspect. When an agent is described as taking responsibility for a *past* action, this involves committing to future actions that express a willingness to make amends for it or to repair damage done by it. But this does not entail being retributively responsible.

David Enoch notes that we may be able to take responsibility for something we have previously done even when we are not to blame for it at all.[26] Consider cases of agent-regret, of the sort described by Bernard Williams, in the face of bad moral luck (e.g., a driver blamelessly hitting a pedestrian).[27] Such cases suggest an ability precisely to exhibit future-directed commitment in relation to actions that were outside of our control, by adopting a willingness to make recompense. Here, the agent is neither retributively accountable nor even attributable (except perhaps to a very weak degree). We would not regard them

24 Pickard, "Responsibility Without Blame: Empathy and the Effective Treatment of Personality Disorder," "Responsibility Without Blame: Therapy, Philosophy, Law," and "Rethinking Justice."

25 Waller, *Against Moral Responsibility*, 104–14.

26 Enoch, "Being Responsible, Taking Responsibility, and Penumbral Agency."

27 Williams, "Moral Luck."

as someone who deserves to suffer in proportion to the harm they have caused, even if we think it is not out of place for them to actively take responsibility by adopting a future-directed commitment to make amends.

Children may also be able to take responsibility for things despite not being retributively blamable if they fail. When a parent asks a child to take responsibility for feeding the hamster, the parent is certainly expecting the child to exhibit futured-directed commitment, but they need not suppose that were the child to fail and were the parent to end up having to feed the pet after all, the child would deserve to suffer retributively. If the parent scolds the child for it, any suffering would naturally be regarded as an instrumental rather than intrinsic good.

Where an agent exhibits future-directed commitment or takes responsibility for something despite not being retributively responsible for it, she must still possess certain abilities: the ability to care about something, to be sensitive to reasons, and to exercise strength of will. This suggests that some degree of attributability *is* required, even if retributive desert is not. But this is true only to a weak degree. Children can exhibit future-directed commitment despite the fact that they do not fully meet the conditions typically associated with attributability, since they lack mature capacities of reason and reflection, do not have a fully developed set of values, and do not reliably succeed in translating their underdeveloped values into choices and actions.

While a child who takes responsibility may perform actions that are attributable to her, she does not count as the sort of agent to whom the conditions of attributability generally apply. In contrast, the blameless but unlucky driver is the sort of agent to whom actions are typically attributable, but this particular action is not attributable to them. The driver has fully developed capacities for reason and reflection and can typically translate their values into choices and actions, but this particular action was accidental and not a true reflection of their values or intentions.

It seems impossible that an agent could exhibit future-directed commitment with respect to something if *neither* the agent nor their relevant actions qualified as attributable to some degree. So some degree of attributability is required for taken responsibility. But neither the child nor the unlucky driver would usually be thought to be fully responsible in the attributability sense. The capacity to take responsibility is distinct, then, from both attributability and retributive desert, even if it is weakly connected to the former.

### *2.3. Skepticism and Incompatibilist Doubts*

For free will skeptics, the capacities associated with attributability are not sufficient for basic retributive desert. And it should be uncontroversial for all sides that the capacities required for taking responsibility or exhibiting

future-directed commitment do not suffice for retributive desert. We need retributive desert in order to justify the retributivist view that it would be an intrinsic good for guilty parties to suffer in proportion to their intentional wrongdoing just because it is deserved.

Skeptical worries typically arise in relation to perceived threats to free will, such as causal determinism, randomness, or pessimism about either one. The argument for regarding these as threats typically draws either on concerns about leeway (whether any agent is capable of choosing otherwise) or else on concerns about ultimate sourcehood (whether any agent is the ultimate source of her own choices) where one or both are taken to be further preconditions for retributive desert.

The arguments for skepticism about retributive desert do not rest on skepticism about whether any agent meets the conditions of attributability—whether, for instance, any agent is acting on purpose or really endorses the desires that motivate her. Even if we are acting on our deeply held values, if these are ultimately explained by factors entirely outside of our control, skeptics argue that this renders punishment purely for the sake of retribution morally suspect. Skepticism about retributive blame neither rests on nor entails skepticism about attributability.

Opponents of skepticism typically suppose that if an agent meets the conditions of attributability, this is sufficient for their meeting the conditions of retributive desert too. Skeptics deny this. Skepticism is usually motivated by some form of incompatibilism with respect to retributive desert (traditionally, seeing it as ruled out by determinism, though skeptics may be concerned that it is ruled out by indeterminism too). But even those who are incompatibilists about retributive desert are typically willing to accept compatibilism about attributability. They simply argue that compatibilism about attributability does not suffice to establish compatibilism about retributive desert.

While our standard desert-entailing practices seem to presuppose that attributability suffices for retributive desert, skeptics endorse revision of these practices and will therefore suppose that their validity cannot be taken for granted: it would be unfair to punish someone just for the sake of retribution if her choices are ultimately fixed by factors outside of her control. This need not entail that there is no difference, say, between actions performed voluntarily and those that are coerced. It just means that acting voluntarily is not sufficient for basic desert. It may be a *necessary* condition, but it cannot be a sufficient one as there are further necessary conditions (i.e., requirements of sourcehood or leeway) that may or may not be met.

While there is some controversy about whether compatibilism about attributability suffices for compatibilism about retributive desert, it should be far

less contentious to say that the sorts of incompatibilist challenge that prompt skepticism about retributive desert entail no corresponding skepticism about taken responsibility, since agents can exhibit future-directed commitment without even fully meeting the conditions of attributability, let alone meeting any further conditions potentially required for retributive desert. Agents like the child or the unlucky driver will not count as retributively blameworthy even by traditional compatibilist standards.

Given these ambiguities, it is important to keep in mind the distinction between different senses of 'responsibility' when assessing the implications of responsibility skepticism. It seems plausible that skepticism about taken responsibility would have terrible implications. But skeptics about retributive desert do not even endorse skepticism about attributability. They certainly do not (and need not) accept skepticism about taken responsibility.

### *2.4. Responsibility and Personhood*

In defending Sen's capacity model of well-being, Caruso emphasizes protecting "the substantive freedom" a person has "to lead the kind of life he or she has reason to value."[28] This use of the word 'freedom' does not entail retributive responsibility. But it does plausibly entail placing moral and practical importance on protecting and encouraging certain capabilities, including those that enable us to exhibit future-directed commitment.

The skills required for taking responsibility are important for a range of reasons that are unconnected to retributive desert. Future-directed commitment is central to our sense of personhood, such that if this is undermined, it is experienced as dehumanizing. Taking responsibility is central to our sense of self-efficacy or our command over our own future behavior. We task children with taking responsibility when we are on the cusp of beginning to treat them as persons. When we do not allow an adult to take responsibility, this is experienced as patronizing. My claim below is that respect for personhood requires respect for the capacities that underscore taken responsibility or future-directed commitment.

## 3. SKEPTICISM AND DEHUMANIZATION

In this section, I will explore a collection of key worries that seem to underscore unease about public health approaches to crime, focusing on four arguments in particular: Peter Strawson's worries about alienating objectivity, Herbert Morris's concern about personhood and the right to be punished, Peter Conrad's

28 Caruso, *Public Health and Safety*, 19.

worries about the medicalization of deviance, and Ken Levy's criticism of Caruso's free will skepticism.[29] While these arguments present diverse considerations, there is a common thread underlying them. They all, in some way or other, suppose that there is something dehumanizing either about responsibility skepticism or about medicine-related approaches to crime—or both. Between them, I think these represent the main categories of argument that motivate unease regarding the public health quarantine model.

I hope to show that there is another common thread between them. They all, to some extent, presuppose that skepticism about retributive desert (and/or medicalized approaches to deviant behavior) must undermine taken responsibility as well. This assumption is essential to motivating the idea that such approaches are impersonal and dehumanizing. I want to argue (a) that we need not accept this assumption, as skepticism about retributive desert does not entail skepticism about taken responsibility, (b) that when we reject it, the public health approach no longer appears dehumanizing, and (c) that Caruso's policies in particular are not dehumanizing in any of the ways suggested by these lines of argument.

### *3.1. The Objective Attitude*

Objections to treating crime as a public health problem often come from a Strawsonian outlook. Strawson argues that skepticism about moral responsibility and a suspension of backward-looking attitudes would be alienating. He equates holding others responsible with seeing them as appropriate targets of reactive attitudes, i.e., "the attitudes and reactions of offended parties and beneficiaries; of such things as gratitude, resentment, forgiveness, love, and hurt feelings."[30] These mark an attitude of "involvement or participation," which we adopt with those we hold morally responsible.[31]

In contrast, when we do not hold a person morally responsible, we adopt a more detached attitude, suspending feelings connected to social demands and expectations. We do not engage with the agent as a person. Rather, we "see him, perhaps, as an object of social policy; as a subject for what, in a wide range of senses, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided."[32] This plausibly captures what is objectionable about denying an agent's responsibility. It is not entirely clear

29 Strawson, "Freedom and Resentment"; Morris, "Persons and Punishment"; Conrad, "Medicine as an Instrument of Social Control"; and Levy, "Let's Not Do Responsibility Skepticism."

30 Strawson, "Freedom and Resentment," 5.

31 Strawson, "Freedom and Resentment," 9.

32 Strawson, "Freedom and Resentment," 9.

what Strawson means by viewing someone "objectively."[33] But the key idea is that being subject to "treatment" or being "managed or handled or cured or trained" involves being treated in an objectionably impersonal manner.

The contrast can be nicely illustrated by reflecting on the plot of Anthony Burgess's novel *A Clockwork Orange* (famously adapted to film by Stanley Kubrick). Alex, a violent criminal, is subjected to two approaches to dealing with convicted criminals, one backward-looking and retributive and the other forward-looking and nonretributive. First, he is placed in a standard prison. It is grotty and unpleasant. He is treated with moral contempt by the prison guards, who enforce a regime of punishment and hold him accountable for his actions. The prison chaplain regularly talks to him and reasons with him. This captures what Strawson calls the "attitude of involvement or participation": Alex is seen as an appropriate target for attitudes like resentment and blame.

Alex is then taken out of this institution and placed in another one. The second institution is a nice, shiny clinic. Here, he is intermittently subjected to a program of conditioning whereby he is forced to watch films of violence while given a drug that makes him feel like he is suffocating (a plot no doubt inspired by real-life examples in which criminals were "conditioned" with drugs like succinylcholine chloride).[34] The goal is to produce an aversion to violence, rendering his future behavior harmless. In this institution, Alex is not blamed or resented, merely, as Strawson would put it, "managed or handled or cured or trained." He is rarely spoken to, since his thoughts are largely irrelevant to what they are doing. This seems a good illustration of Strawson's objective attitude. The fact that this attitude seems dehumanizing is also reflected in the novel, with Alex's anguished plea: "Me, me, me. How about me? Where do I come into all of this? Am I just like some animal or dog?"[35] The worry is that adopting an impersonal attitude across the board would be dehumanizing for us all.

While this "treatment" gives us a clear illustration of an agent being regarded "objectively" in Strawson's sense, it also involves more than just a rejection of retributive blame. Alex's capacity for taken responsibility is also undermined. He is robbed of the power to exhibit future-directed commitment with respect to his own behavior. While Alex's treatment exemplifies the sort of strained objectivity of attitude identified by Strawson, it is not obvious that such strained objectivity is entailed merely by the rejection of retributive desert. We can see this if we think about policies that involve rejecting retributivism

33 On this point, see Tadros, "Treatment and Accountability."

34 We will return to these nonfictional examples in section 4 below.

35 Burgess, *A Clockwork Orange*, 104.

but retaining a strong emphasis on the capacities underlying taken responsibility. This is precisely what Caruso's positive proposals do.

Conditions such as poverty, homelessness, abuse, and domestic violence are factors that can significantly undermine a person's capacity for developing and exercising taken responsibility. They undermine the ability to develop sensitivity to others or to exercise self-mastery. For example, evidence suggests that poverty makes people more impulsive and weak willed and makes it harder to reason about the long-term consequences of one's actions.[36] There is clearly nothing dehumanizing about taking people out of poverty. If anything, being subjected to poverty, homelessness, and abuse is dehumanizing. Tackling these problems strengthens the capacity for taken responsibility rather than weakening it. On my analysis, this reflects a correspondingly strengthened rather than weakened respect for personhood. Similarly, most mental illnesses are commonly acknowledged to be a barrier to the capacities needed for future-directed commitment, so increasing provisions for early treatment and securing free health care are also policies that would strengthen rather than weaken the capacity for taken responsibility. Again, this is hardly dehumanizing.

Caruso suggests that we ought to shift our focus from retribution to "prevention, rehabilitation, and reintegration."[37] Rehabilitation and reintegration essentially require helping offenders to become capable of taking on responsibilities in life outside of prison. This capacity may be best served by being encouraged to take responsibility for one's environment and take on employment roles that better mirror the outside world (opportunities that are typically more limited in prison systems with a heavy emphasis on retribution). Similarly, education improves our capacity to think critically and make informed choices, and hence, education strengthens the abilities that are central to taken responsibility. Again, it seems plausible to think that a lack of access to education rather than increased access is dehumanizing.

Exposure to environmental toxins also reduces one's capacity for taken responsibility, as well as making one more vulnerable to criminality. For example, lead poisoning causes damage to the brain, which affects reasoning ability; those who suffer from it are typically impulsive and less able to exercise self-control. Again, it is hardly dehumanizing to limit the exposure risk of vulnerable populations.

Finally, psychopaths also tend to act impulsively, lack self-control, and be insensitive to others' interests and so are hampered from being able to

36 Mullainathan and Shafir, *Scarcity*; Pepper and Nettle, "The Behavioral Constellation of Deprivation."

37 Caruso, *Public Health and Safety*, 21.

effectively take responsibility. More research into effective interventions and rehabilitation strategies would potentially lead to an increase rather than a decrease in these abilities, and hence this policy also respects personhood.

While Caruso's policy suggestions can hardly be regarded as "objectifying" in the way Strawson takes to be problematic, there remains a worry that the view prevents us from seeing anyone as an apt target for reactive attitudes. Pereboom suggests we could retain *some* reactive attitudes, namely those that are not morally problematic. By substituting resentment or guilt with shock and disappointment or regret at being an agent of a wrong, we may be able to avoid any alienating detachment.[38] But it is not obvious that the reactive attitudes, even those tied to blame, like guilt or resentment, entail *retributivism*. Retributivism is usually understood as the view that the proportionate suffering of a wrongdoer is intrinsically good on the basis that it is deserved. Strawson never mentions retribution in his famous article, so it is not at all clear that Strawson's prime target is skepticism specifically about retributive desert, as opposed to skepticism about weaker forms of responsibility.

In personal relationships, attitudes like resentment do not obviously have retributive implications. If my spouse makes a hurtful comment, feeling resentful may be an unavoidable implication of adopting the attitude of participation. But it is hardly obvious that I must thereby want my spouse to suffer or, even if I do, that I must want this on the basis that I regard such suffering as an *intrinsic good* because it is deserved. In a healthy relationship, we are likely to regard any suffering that comes from expressing resentment as instrumental to fostering greater mutual understanding and empathy rather than seeing it as a means of enacting *retribution*. (The latter goal would be regarded more naturally as a sign of bitter relationship breakdown than as a marker of meaningful engagement.)

There seems to be no central sense, then, in which skepticism about retributivism alone entails the strained objectivity of attitude that Strawson suggests would be so alienating.

### *3.2. The Right to Be Punished*

Morris argues that being retributively punished for our crimes is a *right*; if we are not held responsible as agents, our wrongdoings are inevitably seen as illness, warranting treatment rather than punishment.[39] He gives four reasons for supposing that this is objectionable. First (echoing Alex's lament from *A Clockwork Orange*), if we are not held responsible for our behavior, our status is reduced to that of animals; second, it robs us of the capacity to enjoy any sense of achievement in

38 Pereboom, *Living Without Free Will*, 187–213, and "Free Will, Love, and Anger."

39 Morris, "Persons and Punishment."

relation to what we do; third, "what we receive comes to us through compassion, or through a desire to control us"; and finally, "the logic of cure will push us toward forms of therapy that inevitably involve changes in the person made against his will."[40] This involves being treated like animals or machines—being controlled and manipulated—whether we consent to it or not. Moreover, Morris argues that we have the concept of *cruel punishment* but not that of *cruel treatment* (as opposed to merely painful treatment). Hence, there is no need for procedural safeguards in medicine of the sort we have in the legal system.[41]

The claim that we have a right to exercise taken responsibility would follow more plausibly from these arguments than the claim that we have a right to be retributively punished. It is not at all obvious that a failure to hold others retributively responsible has any of these implications, at least not once we see that this need not involve skepticism about the existence or the moral importance of taken responsibility.

Moreover, Morris seems to endorse a picture of medical ethics according to which it is always permissible for the sake of treatment to bypass an agent's wishes and consent and to inflict manipulative treatments as if we are training an animal or programming a machine. But why should we suppose that this is an ethical approach even to medicine? We now recognize a range of health problems connected specifically to agency—obesity, addiction, eating disorders, depression, anxiety disorders, obsessive compulsive disorders, etc. I would not think much of a doctor who supposed that in treating any of these conditions, it would be okay to treat patients as animals, manipulate them, or inflict treatments on them against their will.

Nor is it obvious that insofar as we regard these as illnesses, a patient who succeeds in getting through a program of recovery is unable to feel any sense of achievement. Programs aimed at treating addiction or obesity commonly involve marking and celebrating achievements, like meeting weight-loss goals or being clean for a year.

And while we may have lacked the concept of cruel treatment at the time Morris was writing, we certainly *do* have this concept now. Many practices that were once considered acceptable (such as forced unsedated electroconvulsive therapy) have since come to be regarded as unduly cruel, and we now recognize a need for legal safeguards.

Historically, problems like addiction and obesity *were* thought to warrant moral contempt rather than treatment. It would be counterintuitive to regard the move away from this attitude and towards a treatment model as a violation

40 Morris, "Persons and Punishment," 486–87.

41 Morris, "Persons and Punishment," 485.

of anyone's rights. Illnesses relating to agency (addiction, depression, compulsion, etc.) typically weaken an agent's ability to effectively assume responsibility for her own behavior. An effective treatment may correspondingly strengthen it. If this capacity is thought to matter morally, this provides a strong moral imperative *not* to inflict entirely manipulative fixes against an agent's will. Such moral imperatives have not always been recognized in the past (as will be further explored below), but a modern medical practitioner is unlikely to suppose that merely classifying something as a medical problem would justify inflicting painful and manipulative treatments on a patient without her consent.

And once again, the abilities that underscore taken responsibility are threatened rather than strengthened by factors such as poverty, exposure to abuse, lack of mental health support and medical care, lack of education, exposure to toxins, etc. So Caruso's policy proposals certainly do not reflect Morris's picture of impersonal or medicalized treatment, since they all aim to strengthen rather than to bypass rational agency.

### 3.3. *Medicalizing Deviance*

There is a longstanding worry about deviant behavior being encompassed within the realm of medical treatment. Thomas Szasz and Nicholas Kittrie each give influential early critiques to this effect, but I am going to focus on Conrad's succinct summary of some of the key dangers associated with the "medicalization" of deviance, which takes into account some of the main lines of arguments developed by earlier theorists.[42]

Conrad identifies at least six categories of problem.[43] First, when a person is seen as ill, Conrad maintains that they are not encouraged to take responsibility. This causes a significant drop in status, as they are essentially tainted with their condition and dependent on those classed as "non-sick." Second, the use of medical language often obscures the value judgments behind medical practices, hiding the moral and political agendas driving public health policy. Third, once something is classed as falling under the remit of medicine, this means it gets taken out of the realm public debate and put into the hands of experts. Fourth, "defining deviant behavior as a medical problem allows certain things to be done that could not otherwise be considered; for example, the body may be cut open or psychoactive medications given."[44] Fifth, once we see something as a medical problem, this pushes us towards an emphasis on the individual,

42 Szasz, *Law, Liberty and Psychiatry*; Kittrie, *The Right to Be Different*; and Conrad, "Medicine as an Instrument of Social Control."

43 Conrad, "Medicine as an Instrument of Social Control," 248–51.

44 Conrad, "Medicine as an Instrument of Social Control," 249–50.

discouraging us from considering the social causes of the problem. Finally, the medicalization of deviant behavior can rob that behavior of political meaning, removing the category of "evil" from our understanding of the world.

Worries about "medicalizing" deviant behavior are not necessarily misplaced. But that term may be given broad or narrow readings. Read narrowly, it encompasses only policies that involve treating an individual's behavior as an illness and seeking to alter it with treatment, ignoring broader societal factors. This *can* be problematic, but the public health model does not count as medicalizing crime on this narrow reading. Read broadly, the term encompasses any strategy that puts something within the broad remit of public health policy. On that reading, the public health model *does* count as medicalizing crime, but this becomes unproblematic.

Is it true that once someone is seen as ill, they are not encouraged to take responsibility and are essentially tainted with their condition? This may be true of some (though hardly all) physical ailments, but there are few courses of treatment for problems like addiction or obesity that do not essentially require an agent to take responsibility and aim to increase the degree to which an agent is able to do this. For example, cognitive behavioral talking therapies aim precisely at enabling agents to exercise future-directed commitment and to more effectively translate their wills into action. Moreover, many public health measures aimed at tackling things like obesity and addiction do *not* essentially taint individuals with their illnesses. Measures for tackling obesity include things like reducing the sugar and fat content in foods, putting clearer and more informative labelling on packages, restricting advertisements for junk food on children's television, adding health and nutrition education to school curricula, removing sweets from next to the checkout in supermarkets, etc.

Relatedly, the idea that individuals must be the sole focus of health interventions is somewhat outdated. For example, Virginia Chang and Nicholas Christakis have examined changes to the entry for 'obesity' in the *Cecil Textbook of Medicine* over a period of one hundred years and found significant shifts over time.[45] In 1927, the focus was entirely on the individual, who was also held personally responsible for overeating. In later editions, the focus shifts towards societal factors that make individuals vulnerable to obesity, such as the wide availability and aggressive marketing of junk foods. By 2000, there is also a focus on the damaging repercussions of blaming individuals for obesity, as this makes them vulnerable to victimization and mental health problems. The picture of public health care in Caruso's model better reflects the trend towards taking a less individualistic approach to health care and addressing societal risk factors.

45 Chang and Christakis, "Medical Modelling of Obesity."

It is perhaps true, as Conrad contends, that the use of medical language *can* obscure value judgments, taking debate out of the public sphere and putting it into the hands of experts, especially where medicalization is construed narrowly. But it is not obvious that all matters of public health policy are like this. Measures that affect the whole public (e.g., sugar and alcohol taxes, low emission zones, smoking bans, pandemic policies, etc.) often spark a great deal of public debate, and the political values in dispute are often transparent—for example, it is often clear that we are weighing personal or commercial freedoms against public safety.

Defining something as a health problem also seems neither necessary nor sufficient for allowing procedures such as cutting open the body or administering psychoactive drugs. Cosmetic surgery involves cutting open the body to "treat" problems that no one regards as illnesses (like small breasts or a crooked nose). And even when something *is* a medical problem, this does not automatically entail that such procedures are justified. We might think that some such procedures are and were *never* justified (e.g., frontal lobotomies and bloodletting). And except in extreme cases, any procedure that goes against the wishes of a patient may be regarded as unjustifiable even if the patient is ill.

Finally, should we worry that Pereboom and Caruso's model removes the category of "evil" from our understanding of the world? Even if we were to regard no one as deserving punishment aimed purely at retribution, we could still class *actions* that aim to harm others as morally wrong and those actions that aim to cause atrocious harms as evil. But the view calls into question whether *people* count as evil. This is a bullet that free will skeptics are typically willing to bite. Those of us who are not skeptics (but are merely doubtful about whether we have adequate epistemic justifications for extensively attributing basic desert to others) need not suppose no one is evil, merely that we should have limited confidence in assessing them as such.

And once again, if we turn specifically to Caruso's policy suggestions, we find that they are not vulnerable to Conrad's worries. They focus predominantly on societal factors, and they aim at increasing an agent's capacity for taken responsibility rather than removing it. (Again, poverty, lack of health care, exposure to toxins, etc. weaken this capacity.) So once again, these policy suggestions do not seem vulnerable to the objection.

### *3.4. Skepticism About Skepticism*

Similar themes recur in Ken Levy's recent critique of Caruso's view.[46] Levy argues that given universal skepticism about desert, "the traditionally recognized excuses—automatism, duress, entrapment, infancy, insanity, involuntary

46 Levy, "Let's Not Do Responsibility Skepticism."

intoxication, mistake of fact, and mistake of law—are suddenly far too limited." Responsibility skeptics "are committed to replacing the recognized excuses with a much broader excuse, a 'universal nonresponsibility' excuse that applies to everybody not because of any cognitive deficiencies or situational constraints but simply because of a metaphysical deficiency: their universal human inability to be genuinely responsible for their crimes."[47]

Echoing Strawson, Morris, and Conrad, Levy argues that the skeptic's position is dehumanizing and threatens human dignity: "Most adults believe that their dignity, which they deeply value, would be severely impaired by others' perception that they are not responsible for their choices and behavior. Such impairment tends to yield devastating effects, including learned helplessness (i.e., fatalistic resignation), diminished cognitive self-efficacy, and lower self-esteem."[48]

He also supposes that Caruso's reasoning would lead to a massive increase in incarceration because it would make sense to preventatively incarcerate those who have committed no crimes so long as they fall into categories that render it likely that they will offend, e.g., having pro-criminal attitudes and values, acquaintances who share these pro-criminal values, personality traits such as hostility and lack of empathy, family problems such as childhood neglect and abuse, low educational attainment, and alcohol or drug problems.[49] Once we stop engaging with someone's behavior as an expression of their own considered and responsible choices, it will inevitably be viewed just like any other impersonal source of danger that might be targeted with risk assessments. The fact that it is a person's own deliberate doing will lose all moral significance.

Levy's claim that Caruso's skepticism about desert entails that all of the traditionally recognized excuses must be thrown out and replaced with a "universal nonresponsibility" excuse fails to take into account the difference between skepticism about *retributive desert* (which free will skeptics are committed to) and skepticism about varying degrees of attributability and taken responsibility (which free will skeptics are not usually committed to). These distinctions would still be incredibly important legally, given that agents can be expected or encouraged to take responsibility only for behavior that is deliberate, informed, uncoerced, etc.

Moreover, we have "learned helplessness" only insofar as we are unable to prospectively take responsibility. It is not obvious that this requires being held *retributively* blameworthy for our past behavior. Again, consider the move away from viewing obesity as a moral failing for which individuals should be

47 Levy, "Let's Not Do Responsibility Skepticism," 3.

48 Levy, "Let's Not Do Responsibility Skepticism," 4.

49 Levy, "Let's Not Do Responsibility Skepticism," 6.

blamed and towards viewing it as a medical problem that calls for effective public health measures. Most such measures presuppose an ability to prospectively take responsibility (as is the case with, e.g., better package labelling and dietary education in school curricula); aim to strengthening agents' ability to exercise strength of will (as is the case with, e.g., cognitive behavioral talking therapies and support groups); or aim to limit exposure to factors that weaken agents' ability to exercise strength of will (as is the case with, e.g., regulations that limit aggressive marketing of junk food).

Throwing away taken responsibility would plausibly produce fatalistic resignation and low self-esteem as Levy supposes. But measures aimed at strengthening agents' capacities for taken responsibility are often at odds with measures aimed at enacting retribution. This is true in relation to crime as well as in relation to traditional health problems: for example, those who wish to make prisons less retributive typically also wish to make them more effective for rehabilitation. In standard UK and US prisons, inmates live in austere cells, are banned from personalizing their spaces, and often have more limited access to mental health support, fewer opportunities for education and training, and fewer opportunities to develop work skills. In contrast, in Norwegian prisons, which are far less focused on punitive measures, inmates are actively encouraged to take responsibility for the spaces they live in, are offered greater opportunities for education and training, and may be given the chance to actively take on work responsibilities mirroring those of outside workplaces.

Punitive systems do not necessarily do anything to encourage inmates either to take more responsibility for their environment and development or to develop skills that will better enable them to take responsibility on their release. If we undermine the capacity for taken responsibility, this really does create fatalistic resignation. But support for retributive blame is often, at best, completely orthogonal to encouraging and enabling greater taken responsibility or, at worse, directly in conflict with it.

Finally, if we suppose that encouraging and enabling taken responsibility is morally important (a stance that we can plausibly adopt consistently with skepticism about retributive blame), then we will have strong reasons not to incarcerate people merely on the basis that they fall into various categories associated with a higher risk of criminality. This obviously robs agents of the opportunity to prospectively take responsibility for their own future behavior by rendering their intentions with respect to their own future behavior irrelevant.[50] There are also other factors that would count against this policy,

50 A related claim by Lemos is that if public safety is the goal, we may have reason to lower the standard of evidence required for conviction from guilt being established beyond reasonable doubt to it merely being likely on the preponderance of evidence. See Lemos,

including some forward-looking considerations that Levy mentions himself, such as the fact that "the resulting rage and terror that would spread throughout the community, would arguably outweigh the public benefit."[51]

Moreover, most of the risk factors themselves are ones that Caruso's policies are directly aimed at tackling (poverty, traumatic childhood experience, addiction, poor access to education, etc.). There is a big difference between policies that aim to prevent people from becoming vulnerable to these risk factors and a policy of incarcerating those who have already been exposed to them. The first strengthens the agents' ability to exhibit future-directed commitment by strengthening the ability to exercise strength of will and make better informed and less impulsive decisions. The second, in contrast, weakens or completely removes this ability. While it is dehumanizing to undermine an agent's ability to take responsibility, it is not dehumanizing to withhold retributive blame, and the second stance does not entail the first.

All the theorists discussed in this section seem to share an assumption: that viewing crime as a public health problem and/or rejecting retributive principles entails that we must also be blind to the moral importance of the sorts of abilities that underlie taken responsibility. I have argued that there is no such entailment and that without this entailment, accusations that this approach would justify impersonal or dehumanizing treatment are baseless. If so, we might wonder where the persistent worry about this comes from.

Levy acknowledges that Caruso provides moral reasons why skepticism about retributive responsibility would not justify locking up great swathes of the population who have committed no crime, but he nonetheless claims that "once culpability was abandoned, such reasons would be inadequate barriers to punishment for suspected dangerousness. Given human nature, at least humans' track record for the past few centuries, it is quite likely that even a morally advanced responsibility-skeptical society would simply override these moral principles by filling the space previously occupied by culpability with a much more robust, single-minded concern for public safety."[52] In fact, suspicion that treating crime as a public health problem would have dehumanizing implications is certainly encouraged by the actual history of projects aimed

"A Moral/Pragmatic Defense of Just Deserts Responsibility" and *Free Will's Value*, 149–56. This would not leave huge segments of the population powerless over their lives (as per Levy's suggestion), but it would increase the risk of having our capacities for taken responsibility undermined. If this is a serious harm in itself, then it is not clear that the safety gains will be worth the increased risk, especially if we want to promote not mere safety but also the substantive freedom to lead the kind of life we have reason to value.

51 Levy, "Let's Not Do Responsibility Skepticism," 6.

52 Levy, "Let's Not Do Responsibility Skepticism," 6.

at treating crime as a public health problem. Levy is justly discouraged by our "track record for the past few centuries." It may be this history that continues to provoke unease. We will turn to this point next.

## 4. CRIME, MEDICINE, AND ETHICS

### *4.1. The Dark History of the Association of Medicine and Crime*

The troubling *Clockwork Orange* picture of what might be entailed by "treating" criminality is not restricted to fiction. Ralph Schwitzgebel documents a host of behavior modification techniques that have been used to treat offenders, including methods that draw on classical and operant conditioning.[53] Some rely on positive reinforcement through token economies or tier systems. Notably, however, some rely on various forms of negative reinforcement, including "aversive suppression" techniques involving the administration of electric shocks or the use of succinylcholine chloride, described as "a curare-like drug that rapidly produces complete paralysis of the skeletal muscles, including those which control respiration," resulting in "great fright about being unable to breathe and a fear of suffocation."[54] Such negative reinforcement techniques were used to treat a great many "crimes," including homosexuality, transvestitism, and fetishism.

Psychiatry has been used throughout history as an instrument of social control, from Samuel Cartright's notorious diagnosis of *drapetomania* (the supposed "disorder" of slaves who wished to escape slavery) to the psychiatric internment of Soviet dissidents in the USSR.[55] Moran notes that medicalizing criminality has been associated with numerous morally and scientifically dubious interventions that aim to identify the "born criminal"—a project frequently steeped in racism and classism.[56] Dubious historical attempts to give medical explanations of crime include physiognomy and phrenology, both pioneered in the early nineteenth century. The former sought to diagnose criminality through features of the face, while the latter sought to diagnose criminality through the shape of the skull.[57] Some historical attempts to think

53 Schwitzgebel, *Development and Legal Regulation of Coercive Behavior Modification Techniques with Offenders*.

54 Schwitzgebel, *Development and Legal Regulation of Coercive Behavior Modification Techniques with Offenders*, 10.

55 Cartwright, "Report on the Diseases and Physical Peculiarities of the Negro Race"; and Fareone, "Psychiatry and Political Repression in the Soviet Union."

56 Moran, "The Search for the Born Criminal and the Medical Control of Criminality."

57 Lavater, *Essays on Physiognomy*; and Spurzheim, *The Physiognomical System of Drs. Gall and Spurzheim*.

of criminology in biological terms are absurd to the point of comedy, such as Richard Dugdale's 1870s inquiry into whether pauperism (alongside other elements of "degeneracy" and "criminality") might be hereditary.[58]

These projects have often had racist motivations. Earnest Hooton's study of the link between biology and crime involved comparing prison populations to those outside of prison and concluding that some races were inherently criminal and should be counted as inferior. He began with overtly racist commitments and assumed without question that the process via which some people ended up in prison in 1930s America was neutral and free of bias.[59]

The goal of reducing criminality has also been implicated in the liberal use of involuntary sterilization, particularly in the United States. Targeted "crimes" or "sins" include homosexuality and masturbation. Forced sterilization was associated with racism and eugenics, alongside more well-meaning goals.[60] One of the earliest explicit statements of the claim that "violence is a public health problem" is from Vernon Mark and Frank Erwin.[61] They, along with William Sweet, proposed, initially in response to urban riots, that psychosurgery should be considered for use on large segments of the population as a means of preventing crime.[62] There is some justice in Peter Breggin's description of such proposals as a sort of "psychiatric totalitarianism."[63]

Worries about the "psychiatric totalitarian" potential of associating crime with health are thus not unfounded. Medicine has often been a mask for social control and has been associated with appalling policies and interventions, often inflicted without consent on those deviating from norms. As Emily McTernan argues, this sort of history ought to provoke some moral concern, particularly about certain sorts of medical interventions for deviance such as "neurointerventions."[64]

### *4.2. The Dark History of Medicine Itself*

It is evident even from this very brief summary that the history here is troubling. Nonetheless, I want to suggest that what is troubling about it actually has very little to do with treating crime as a public health problem. The trouble arises from a morally suspect approach to medicine more generally. Many of

58 Dugdale, *The Jukes.*

59 Hooton, *Crime and the Man.*

60 See Largent, *Breeding Contempt*, especially 11–38.

61 Mark and Ervin, *Violence and the Brain*, 160.

62 Mark, Erwin, and Sweet, "Role of Brain Disease in Riots and Urban Violence."

63 Breggin, "Psychosurgery for Political Purposes," 847.

64 McTernan, "Those Who Forget the Past."

the treatments for illnesses that most of us accept ought to be counted within the realm of medicine also have a dark history. Perhaps the problem is not that such illnesses (along with criminality) are regarded as matters of public health but that governments and experts have often exercised poor moral judgment about medicine.

Negative reinforcement in the form of aversive stimuli such as electric shocks and paralyzing drugs has been used to treat not only those behaviors regarded as criminal but also true medical conditions. For example, emetic drugs and succinylcholine chloride have been used in conditioning treatments for alcoholism.[65] Forced and unanesthetized electroconvulsive therapy (ECT) and psychosurgeries, such as frontal lobotomies, have been used to treat mental health problems, including depression, anxiety, addiction, and schizophrenia. For a long time, it was rare to seek the consent of patients at all.[66] Even after laws were introduced requiring informed consent for psychosurgery (which was as late as the 1950s), the extent to which patients were able to count as meaningfully consenting is contentious.[67] Forced sterilization was used to treat various mental health conditions. For example, hysterectomies were used to treat "women's hysteria," which could include psychiatric conditions and epilepsy.[68] Forced sterilizations were also seen as appropriate for preventing the spread of "drunkenness."[69] Some authors advocated castration to stop the breeding of "imbeciles and paupers."[70]

Unsurprisingly, some critics of the use of medical approaches in relation to crime are also skeptical about the treatment of mental illness across the board, arguing that the mind should be entirely outside the sphere of health care. Szasz has written critiques of both the use of medical methods in relation to crime and the inclusion of mental health within the realm of medicine.[71] This stance on mental health is rarely regarded as plausible. Moreover, it seems to misidentify the source of the moral concern. When we contemplate what is wrong with forcing hysterectomies on nonconsenting women as a treatment for epilepsy, the thing that troubles us is not that epilepsy is being erroneously regarded as a medical condition. Epilepsy plausibly *is* a medical condition. Clearly, that

65 Schwitzgebel, *Development and Legal Regulation of Coercive Behavior Modification Techniques with Offenders*, 14.

66 Ottosson and Fink, *Ethics in Electroconvulsive Therapy*, 33–48.

67 Raz, *The Lobotomy Letters*, 69–100.

68 Largent, *Breeding Contempt*, 18–19.

69 Largent, *Breeding Contempt*, 26.

70 Baldwin, "Whipping and Castration as Punishments for Crime," 382.

71 Szasz, *Law, Liberty and Psychiatry* and *Ideology and Insanity.*

alone does not entail that "treating" it with forced hysterectomies is justifiable, either morally or medically.

The danger of patients having treatments inflicted on them without informed consent is something that has increasingly come under scrutiny in medicine. Elizabeth Symonds argues that forced psychotropic treatments should be regarded as a "cruel and unusual punishment" both in penal and in nonpenal settings such as psychiatric units.[72] We now accept, *contra* Morris, that "treatment" can be cruel in addition to merely being painful. Many previously commonplace practices in psychiatry are rejected now precisely on this basis, and we recognize (also *contra* Morris) the need for procedural safeguards.

It is also far from obvious that the dangers that arise in relation to medical treatment of psychiatric ailments are fundamentally different from those that arise in relation to treatment of physical health conditions. There was no clear notion of informed consent in *any* area of medicine until the 1950s, and there is evidence that before that point, while some practitioners consulted patients on whether they wanted to undergo procedures, others regularly failed to.[73] Across the board, history has been patchy with respect to allowing patients to exercise agency and autonomy over the treatments they undergo. Across all areas of medicine, this has improved through increased moral scrutiny and legislation.

But there is probably nothing inherently special about medicine here. If we closely examine the history of marriage, religious organizations, educational establishments, families, workplaces, military organizations, or virtually any other human social institution, we find similar patterns: frequent abuses of power and exploitation of the vulnerable with little regard for individual autonomy or consent—until increased moral scrutiny brings about legislative changes. While medicine has often been an instrument of social control, so has almost everything.

## 5. Ethics, Responsibility, and Public Health

If almost everything has a dark history, this has some implications for how we ought to respond to the dark history of the association between crime and public health. Instead of seeking to stop anything from falling within the remit of public health, we should instead ask why public health initiatives have often been unethical and corrupt. The answer is not, I suspect, because such initiatives are not governed by principles of retributive blame. After all, retributive punishment plainly has an even darker history. Forced unanesthetized ECT is

72 Symonds, "Mental Patients' Rights to Refuse Drugs."

73 Faden, et al., *A History and Theory of Informed Consent*, 53–85.

probably *less* dehumanizing than being publicly disemboweled, burned alive, or crucified.

It is also a false dichotomy to suppose that if we do not view something as a matter of retributive blame, we must view it as entirely outside the realm of taken responsibility. It is false that if we do not treat alcoholism as a moral failing that ought to be punished, then we must instead support forcing alcoholics into programs of aversive conditioning and inflicting horrors on them like electrocution or succinylcholine chloride.

From an ethical perspective, there is a critical difference between emphasizing public policy measures that reduce the risk to vulnerable groups of developing certain problems (whether it be criminality, obesity, addictions, heart disease, seasonal flu, or whatever) and policy measures that needlessly weaken agents' abilities to assume control over their own future behaviors. The reason why Caruso does not move from a lack of retributive blame directly to an endorsement of mass incarceration for those who fall into various risk categories or to a program of coercive drugging or conditioning of offenders is because skepticism about retributive blame does not entail that the capacity of agents to exercise future-directed commitment with respect to their own behavior is no longer a valid moral concern. Nor does it entail that consent is never required for any effective intervention. If we think that these *are* valid moral concerns, then we will have every reason to class these strategies as unethical methods of both crime prevention *and* medical treatment.

## 6. CONCLUSION

This paper has sought to challenge a common source of uneasiness about treating crime as a public health problem. It is an uneasiness that derives from a history of medicalizing crime that is indeed ethically problematic. The worry is that once we put crime within the remit of medicine, we must endorse impersonal, manipulative, and dehumanizing measures of tackling crime.

The mistake, I maintain, does not consist in our putting crime within the broad remit of public health but in supposing that impersonal, manipulative, and dehumanizing measures would become morally acceptable the moment that crime (or anything else) is placed within the boundaries of public health. The problem is that we have often had a lax moral approach to health measures. Critiques that continue to be highly influential, such as Strawson's and Morris's, emerged in the 1960s, after several decades, if not centuries, in which standard practices for dealing with mental health problems included measures that we now view as shockingly unethical and inhumane. Perhaps at that time, it seemed obvious, as Morris contended, that we could treat those with illnesses

like animals, ignore their consent, inflict cruel treatments, etc., but we should not have thought this was justifiable in the name of medicine then, and we need not suppose that this is an acceptable approach to medicine now.

One major virtue of Caruso's policy suggestions is that they reflect an ethically sensitive picture of what good public health policy should look like. Public health measures across the board should adhere to defensible ethical standards. While many of these standards are tied to some recognition of the moral importance of protecting and cultivating the capacities that underlie taken responsibility or future-directed commitment, they are not tied essentially to retributive blame.[74]

*University of Warwick*
*nadine.elzein@warwick.ac.uk*

## REFERENCES

Baldwin, S. "Whipping and Castration as Punishments for Crime." *Yale Law Journal* 8, no. 9 (1899): 371–86.

Breggin, Peter R. "Psychosurgery for Political Purposes." *Duquesne Law Review* 13, no. 4 (1975): 841–62.

Burgess, Anthony. *A Clockwork Orange*. Banned Books series. Paperview, 1962.

Cartwright, Samuel A. "Report on the Diseases and Physical Peculiarities of the Negro Race." *New Orleans Medical and Surgical Journal* 7 (1851): 691–715.

Caruso, Gregg D. "Free Will Skepticism and Criminal Behavior: A Public Health-Quarantine Model." *Southwest Philosophy Review* 32, no. 1 (2016): 25–48.

———. *Public Health and Safety: The Social Determinants of Health and Criminal Behavior*. ResearchLinks Books, 2017.

———. *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice*. Cambridge University Press, 2021.

Caruso, Gregg D., and Derk Pereboom. *Moral Responsibility Reconsidered*. Cambridge University Press, 2022.

———. "A Non-punitive Alternative to Retributive Punishment." In *The Routledge Handbook of the Philosophy and Science of Punishment*, edited by Farah Focquaert, Elizabeth Shaw, and Bruce N. Waller. Routledge, 2020.

Chang, Virginia W., and Nicholas A. Christakis. "Medical Modelling of Obesity:

74 I am grateful to Dr. Tuomas K. Pernu, Professor Victor Tadros, Dr. Robyn Repko Waller, anonymous reviewers, and *JESP*'s anonymous associate editor for incredibly helpful comments on earlier drafts of this paper.

A Transition from Action to Experience in a 20th-Century American Medical Textbook." *Sociology of Health and Illness* 24, no. 2 (2002): 151–77.

Conrad, Peter. "Medicine as an Instrument of Social Control: Consequences for Society." In *Deviance and Medicalization: From Badness to Sickness*, edited by Peter Conrad and Joseph W. Schneider. Temple University Press, 1992.

De Ruyter, Doret. "The Virtue of Taking Responsibility." *Educational Philosophy and Theory* 34, no. 1 (2002): 25–35.

Dugdale, R. L. *The Jukes: A Study in Crime, Pauperism, Disease, and Heredity*. G. P. Putnam's Sons, 1877.

Elzein, Nadine. "Undetermined Choices, Luck and the Enhancement Problem." *Erkenntnis* 88, no. 7 (2023): 2827–46.

Enoch, David. "Being Responsible, Taking Responsibility, and Penumbral Agency." In *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams*, edited by Ulrike Heuer and Gerald Lang. Oxford University Press, 2012.

Faden, Ruth R., Tom L. Beauchamp, and Nancy M. P. King. *A History and Theory of Informed Consent*. Oxford University Press, 1986.

Faraone, Stephen. "Psychiatry and Political Repression in the Soviet Union." *American Psychologist* 37, no. 10 (1982): 1105–12.

Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, 1998.

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

Gaden, Gerry. "Rehabilitating Responsibility." *Journal of Philosophy of Education* 24, no. 1 (1990): 27–39.

Hobbes, Thomas. "Of Liberty and Necessity." In *Hobbes and Bramhall on Liberty and Necessity*, edited by Vere Chappell. Cambridge University Press, 1999.

Hooton, Earnest Albert. *Crime and the Man*. Harvard University Press, 1939.

Hume, David. *An Enquiry Concerning Human Understanding*. Edited by Tom L. Beauchamp. Oxford University Press, 1748/1999.

———. *A Treatise of Human Nature*. Edited by David Fate Norton and Mary J. Norton. Oxford University Press, 1740/2000.

Kelly, Erin I. "Criminal Justice Without Retribution." *Journal of Philosophy* 106, no. 8 (2009): 440–62.

———. *The Limits of Blame: Rethinking Punishment and Responsibility*. Harvard University Press, 2018.

Kittrie, Nicholas N. *The Right to Be Different: Deviance and Enforced Therapy*. John Hopkins Press, 1971.

Largent, Mark A. *Breeding Contempt: The History of Coerced Sterilization in the United States*. Rutgers University Press, 2011.

Lavater, Johann Caspar. *Essays on Physiognomy: For the Promotion of the Knowledge and the Love of Mankind*. Vol. 1. C. Whittingham, 1804.

Lemos, John. *Free Will's Value: Criminal Justice, Pride and Love*. Routledge, 2023.

———. "A Moral/Pragmatic Defense of Just Deserts Responsibility." *Journal of Information Ethics* 28, no. 1 (2019): 73–94.

Levy, Ken M. "Let's Not Do Responsibility Skepticism." *Journal of Applied Philosophy* 40, no. 3 (2023): 1–15.

Mark, V. H., and F. R. Ervin. *Violence and the Brain.* Harper and Row, 1970.

Mark, V. H., W. H. Sweet, and F. R. Ervin. "The Role of Brain Disease in Riots and Urban Violence." *Journal of the American Medical Association* 203, no. 5 (1968): 368–69.

McTernan, Emily. "Those Who Forget the Past: An Ethical Challenge from the History of Treating Deviance." In *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice, Engaging Philosophy*, edited by David Birks and Thomas Douglas. Oxford University Press, 2018.

Moran, Richard. "The Search for the Born Criminal and the Medical Control of Criminality." In *Deviance and Medicalization: From Badness to Sickness*, edited by Peter Conrad and Joseph W. Schneider. Temple University Press, 1992.

Morris, Herbert. "Persons and Punishment." *Monist* 54, no. 2 (1968): 475–501.

Morse, Stephen J. "Compatibilist Criminal Law." In *The Future of Punishment*, edited by Thomas A. Nadelhoffer. Oxford University Press, 2013.

Mullainathan, Sendhil, and Eldar Shafir. *Scarcity: Why Having Too Little Means So Much*. Macmillan, 2013.

Ottosson, Jan-Otto, and Max Fink. *Ethics in Electroconvulsive Therapy*. Brunner-Routledge, 2012.

Pepper, Gillian V., and Daniel Nettle. "The Behavioral Constellation of Deprivation: Causes and Consequences." *Behavioral and Brain Sciences* 40 (2017): e314. https://doi.org/10.1017/S0140525X1600234X.

Pereboom, Derk. *Free Will, Agency, and Meaning in Life*. Oxford University Press, 2014.

———. "Free Will, Love, and Anger." *Ideas y valores* 58, no. 141 (2009):169–89.

———. "Incapacitation, Reintegration, and Limited General Deterrence." *Neuroethics* 13, no. 1 (2020): 87–97.

———. *Living Without Free Will*. Cambridge University Press, 2006.

Pereboom, Derk, and Gregg D. Caruso. "Hard-Incompatibilist Existentialism: Neuroscience, Punishment, and Meaning in Life." In *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, edited by Gregg D. Caruso and Owen Flanagan. Oxford University Press, 2018.

Pickard, Hannah. "Responsibility Without Blame: Empathy and the Effective Treatment of Personality Disorder." *Philosophy, Psychiatry, and Psychology*

18, no. 3 (2011): 209–24.
———. "Responsibility Without Blame: Therapy, Philosophy, Law." *Prison Service Journal* 213 (2014): 10–16.
———. "Rethinking Justice: The Clinical Model of Responsibility Without Blame." *Howard League for Penal Reform: Early Academics Network Bulletin* 36 (2018): 4–10.
Raz, Mical. *The Lobotomy Letters: The Making of American Psychosurgery*. University of Rochester Press, 2013.
Scanlon, T. M. "Interpreting Blame." In *Blame: Its Nature and Norms*, edited by Justin D. Coates and Neal A. Tognazzini. Oxford University Press, 2012.
Schwitzgebel, Ralph K. *Development and Legal Regulation of Coercive Behavior Modification Techniques with Offenders*. National Institute of Mental Health, Center for Studies of Crime and Delinquency, 1971.
Sen, Amartya. *Development as Freedom*. Oxford University Press, 1999.
Spurzheim, J. G. *The Physiognomical System of Drs. Gall and Spurzheim*. 2nd ed. Baldwin, Cradock and Joy, 1815.
Strawson, Peter F. "Freedom and Resentment." *Proceedings of the British Academy* 48 (1962): 187–211.
Symonds, Elizabeth. "Mental Patients' Rights to Refuse Drugs: Involuntary Medication as Cruel and Unusual Punishment." *Hastings Constitutional Law Quarterly* 7 (1979): 701–38.
Szasz, Thomas S. *Ideology and Insanity: Essays on the Psychiatric Dehumanization of Man*. Syracuse University Press, 1991.
———. *Law, Liberty and Psychiatry: An Inquiry into the Social Uses of Mental Health Practices.* Macmillan, 1963.
Tadros, Victor. "Treatment and Accountability." In *Lanson Lectures in Bioethics (2016–2022): Assisted Suicide, Responsibility, and Pandemic Ethics*, edited by Hon-Lam Li. Palgrave Macmillan, 2024.
Vargas, Manuel. "The Trouble with Tracing." *Midwest Studies in Philosophy* 29, no. 1 (2005): 269–91.
Waller, Bruce. *Against Moral Responsibility.* Massachusetts Institute of Technology Press, 2011.
———. "Virtue Unrewarded: Morality Without Moral Responsibility." *Philosophia* 31 (2004): 427–47.
Watson, Gary. "Free Agency." *Journal of Philosophy* 72, no. 8 (1975): 205–20.
———. "Two Faces of Responsibility." *Philosophical Topics* 24, no. 2 (1996): 227–48.
Williams, Bernard. "Moral Luck." *Aristotelian Society Supplementary Volume* 50 (1976): 115–35.
Wolf, Susan. *Freedom Within Reason.* Oxford University Press, 1990.

*Journal of Ethics and Social Philosophy*
VOL. 29, NO. 2 · JANUARY 2025
*https://doi.org/10.26556/jesp.v29i2.3316*


# IT'S ONLY NATURAL!

## MORAL PROGRESS THROUGH DENATURALIZATION

*Charlie Blunden*

MORAL PROGRESS occurs when things change for the better, morally speaking. Questions of moral progress have recently been receiving increasing interest from philosophers.[1] But how does moral progress happen? This question concerns the *causality of moral progress.*[2] In this paper, I seek to advance the discussion on a potential cause of moral progress that I will refer to as *denaturalization*.[3]

Denaturalization has been investigated by several philosophers in the moral progress literature, most notably Nigel Pleasants, Julia Hermann, Dale Jamieson, and Elizabeth Anderson.[4] The idea is that moral progress can be facilitated by people coming to have a more accurate understanding of the extent to which their institutions are natural or necessary. Proponents of denaturalization as a cause of moral progress argue that progressive moral change is often blocked by a false understanding on behalf of relevant social actors that their current institutional setup is in some way "natural and indispensable."[5] These beliefs

1 For an overview, see Sauer et al., "Moral Progress."

2 Extant theories include that moral progress is caused by greater knowledge of the moral facts (see Huemer, "A Liberal Realist Answer to Debunking Skeptics"); by adaptively plastic psychological mechanisms that respond to increased material security (see Buchanan and Powell, *The Evolution of Moral Progress*, ch. 6); or by the exercise of moral consistency reasoning under favorable social conditions (see Kumar and Campbell, *A Better Ape*).

3 I borrow the term *denaturalization* from Jaeggi, *Critique of Forms of Life*, 8, though I make no claim to be using the term in her sense. Rather, I am using the term to refer to a proposed cause of moral progress discussed by several philosophers in the moral progress literature, described below.

4 See Pleasants, "The Structure of Moral Revolutions" and "Moral Argument Is Not Enough"; Anderson, "Social Movements, Experiments in Living, and Moral Progress"; Jamieson, "Slavery, Carbon, and Moral Progress"; and Hermann, "The Dynamics of Moral Progress."

5 Hermann, "The Dynamics of Moral Progress," 305. See also Jamieson, "Slavery, Carbon, and Moral Progress," 177–80; Pleasants, "Moral Argument Is Not Enough," 166; and Anderson, "Social Movements, Experiments in Living, and Moral Progress," 16.

can often be a significant impediment to changes away from an unjust status quo, and undermining them can be a significant cause of moral progress, as the unjust status quo is then left with no "veneer of naturalization" to hide behind.[6]

The paradigm case that denaturalization is meant to explain is the successful abolitionist movement in nineteenth-century Britain, and I will explore this case in more depth in the first section. Denaturalization has also been implicated in other past or potential instances of moral progress. Hermann points out that appeals to naturalness have played a role in defending practices of discrimination against homosexuality and the oppression of women, which may imply that, to the extent that these practices have been undermined, denaturalization has played a role.[7] Proponents of denaturalization have also suggested that it may have a role to play in moving away from a carbon-intensive economy or in challenging the view that "there is no plausible alternative to wage labor and the market economy" so that an alternative and morally preferable economic system, if one is indeed possible, can be adopted.[8]

The current literature on denaturalization as an explanation of moral progress contains some vagueness about what denaturalization is and how it works, which makes it difficult to work out: what exactly denaturalization is; what empirical presuppositions need to be correct for denaturalization to be a psychologically realistic account of how moral progress happens; and whether and under what conditions denaturalization might lead to moral progress. Thus, my main aim is to develop, using the existing literature as a guide, a more detailed and explicit account of what denaturalization is and how it might work so that the aforementioned points of unclarity can be made clearer.

This paper has four sections. In the first section, I specify denaturalization by clarifying the different interpretations one could have of claims that a given practice or institution is natural or necessary. I argue that the interpretation most compatible with the existing literature is that claims of naturalness or necessity are claims about the *costs* of getting rid of existing institutions and moving to an alternative. In the second and third sections, I develop what I call a *costs account of denaturalization*. In the second section, I explicate a general framework, using recent advances in philosophical understandings of conventionality, which enables us to understand claims of naturalness and necessity as

6 Hermann, "The Dynamics of Moral Progress," 307; and Jamieson, "Slavery, Carbon, and Moral Progress," 180.

7 Hermann, "The Dynamics of Moral Progress," 307.

8 On the potential role of denaturalization in moving away from a carbon-intensive economy, see Jamieson, "Slavery, Carbon, and Moral Progress," 177–78. On its potential role in overcoming the notion that there is no alternative to wage labor and a market economy, see Pleasants, "Moral Argument Is Not Enough," 176–77.

claims about the costs of abandoning status quo institutions and to understand how these claims can be mistaken in degrees. In the third section, I present a brief case for the psychological realism of this account of denaturalization. I suggest that the costs account has some claim to being psychologically realistic, while also highlighting the limits of this claim and outlining the kinds of empirical evidence that proponents of denaturalization need for a convincing account of the psychological realism of denaturalization as a cause of moral progress. Fourth, with the more detailed costs account of denaturalization in hand, I investigate whether and under what conditions denaturalization can lead to moral progress.

## 1. DISAMBIGUATING DENATURALIZATION

In this section, I will introduce the idea of denaturalization as it has previously been discussed in the literature, clarify some possible interpretations of denaturalization, and make explicit which interpretation I am adopting. To introduce denaturalization and clarify the interpretations of it that one could hold, I will first consider in greater depth the paradigm example of denaturalization: British abolitionism in the nineteenth century.[9]

Historically, slavery was widely seen as a natural practice without alternative. As the historian Seymour Drescher documents, for most of recorded human history, slavery has been a ubiquitous institution, viewed as "part of the natural order," and the presence of slavery was so taken for granted that its existence "set limits on how a social order could be imagined."[10] Even by the time of the eighteenth century, estimates put the number of unfree laborers (enslaved persons, serfs, and people otherwise in bondage) at 95 percent of the global population.[11] People throughout history have recognized that enslaved people suffer greatly. Bernard Williams observes that, in ancient Greece, people who were slaveowners or otherwise benefited from slavery nonetheless "granted that [slavery] was intensely unpleasant for the slaves."[12] In the same vein, Thomas Haskell emphasizes that "the suffering of slaves had long

9 I am focusing on the case of British abolition because this is the case most commonly discussed by proponents of denaturalization. In doing so, I am not claiming that abolitionist movements in other countries were less important or less instrumental in eventually ending legalized slavery worldwide. Thanks to an anonymous reviewer for pointing out this potential unclarity.

10 Drescher, *Abolition*, ix. The ubiquity of slavery is also made apparent in Holslag, *A Political History of the World*, especially 540, 551, 555–56.

11 Drescher, *The Mighty Experiment*, 14.

12 Williams, *Shame and Necessity*, 109.

been recognized" before the eighteenth century, but this recognition had not previously led to "active opposition to the institution of slavery."[13] In addition, articulated arguments against slavery go back at least to the time of Aristotle.[14] Thus, prior to abolition, the suffering of enslaved people was recognized, and arguments that slavery was immoral had long been articulated, but these factors did not lead to any sustained efforts to abolish slavery.

Why was this the case? Proponents of denaturalization argue that people often thought that slavery was a necessary economic institution without which it was impossible to produce a social surplus and that this perception made abolishing slavery an unacceptable idea.[15] Bolstering this claim is the observation that moral arguments *in favor* of slavery (often referring to the purported moral responsibility of slave owners and/or the racial inferiority of enslaved people) were quite uncommon until the mid-eighteenth century.[16] Pleasants argues that this lack of positive justifications for slavery until very late in the institution's history is indicative of the fact that for the majority of that history, it was simply taken for granted: for most of its existence, slavery was seen as a "natural, necessary, and inevitable feature of the social world."[17]

In the eighteenth century, wage labor became increasingly widespread. This provided a salient alternative institution to slavery: after all, it was obvious that a substantial social surplus could be produced via the institution of wage labor. This "cracked" the "veneer of naturalization" that had previously attached to the institution of slavery.[18] Prior to the British abolition of slavery in 1833, specific experiments with wage labor had been trialed in former slave plantations in Barbados in the 1780s and 1790s; in Trinidad in 1806 and subsequently in 1812–15 when American former enslaved persons settled there; in Sierra Leone from 1792 onwards; and most notably, in Venezuela in the 1830s, where the number of enslaved persons had been drastically reduced due to legislated freedom

13 Haskell, "Convention and Hegemonic Interest in the Debate over Antislavery," 848.

14 Cambiano, "Aristotle and the Anonymous Opponents of Slavery."

15 Anderson, "Social Movements, Experiments in Living, and Moral Progress," 14–15; Williams, *Shame and Necessity*, 111–13, 124–25; Pleasants, "Moral Argument Is Not Enough" and "The Structure of Moral Revolutions"; and Hermann, "The Dynamics of Moral Progress."

16 Brown, *Moral Capital*, 35–36, 52; and Jamieson, "Slavery, Carbon, and Moral Progress."

17 Pleasants, "Moral Argument Is Not Enough," 166; see also 165n4. I will explore further in section 4 how instances of denaturalization can lead to the emergence of ideological justifications for continued injustice.

18 Hermann, "The Dynamics of Moral Progress," 307; and Jamieson, "Slavery, Carbon, and Moral Progress," 180.

at birth, and agricultural output had been flourishing.[19] These instances of wage labor replacing slave labor were appealed to in parliamentary debates on whether or not to abolish slavery in the British Empire. Proponents touted the proposed Slavery Abolition Act as a "mighty experiment" in free labor that would have morally weighty consequences for as yet unborn subjects of the British Empire and for the "welfare of millions of slaves in foreign colonies."[20] Opponents disagreed, calling it "a procedure with disproportionate social risks—a 'mere,' 'hasty,' or 'dangerous' experiment."[21] More generally, British abolitionists, though often respected members of the bourgeoisie (and thus deeply involved in the wage labor system), were often "denounced as quixotic knights-errant, as pious charlatans all too happy to ruin the empire with costly and disastrous experiments in social engineering."[22] The Slavery Abolition Act was passed in 1833, although enslaved people in the British Empire were not in fact freed until 1838 when campaigns to end the transitionary apprenticeships that continued to bind former enslaved persons to their former masters were successful.[23] For proponents of denaturalization, the morally transformative abolition of slavery came about, at least in significant part, because the emergence of widespread wage labor denaturalized the institution of slavery and thus enabled moral criticism of slavery to become effective and led to the abolition of the practice.[24]

Before moving on to consider how we might understand the notion of naturalness and necessity, I will consider a reasonable response to this historical narrative of British abolition: Why does it focus so much on the perceptions and actions of slaveholders and others who benefitted from or tolerated slavery rather than focusing on the perceptions and actions of enslaved people? After all, it is plausible that enslaved people have always known that slavery is wrong and have always been motivated to overthrow the institution. The issue is that due to their position of extreme disadvantage relative to their enslavers,

19 Drescher, *The Mighty Experiment*, 91–94, 108–20.

20 Drescher, *The Mighty Experiment*, 123; and Anderson, "Social Movements, Experiments in Living, and Moral Progress," 17–18.

21 Drescher, *The Mighty Experiment*, 124.

22 On abolitionists often being members of the bourgeoisie, see Haskell, "Capitalism and the Origins of the Humanitarian Sensibility," 341–46. See also Davis, *The Problem of Slavery in the Age of Revolution*, 81–82. On the denouncements that they were subject to, see Brown, *Moral Capital*, 10.

23 Drescher, *Abolition*, 264.

24 Anderson, "Social Movements, Experiments in Living, and Moral Progress," 15–24; Pleasants, "Moral Argument Is Not Enough," 175–76; and Hermann, "The Dynamics of Moral Progress," 306–7.

enslaved persons have almost never successfully overthrown slavery through their own actions—with the very notable exception of the Haitian Revolution.[25] For instance, around the time of the Slavery Abolition Act being passed in Britain, the "Baptist War" erupted in Jamaica. It was the largest slave rebellion in the history of the British Empire, involving one-fifth of the population of enslaved people on the island (nearly sixty thousand people). However, this uprising lasted only eleven days, from December 25, 1831, to January 5, 1832, due to the limited power of enslaved people to resist heavily armed colonial militias.[26] As such, an explanation for the abolition of slavery, in the British case and likely in other cases besides, must extend beyond the agency of enslaved people to include the agency of the people who were not enslaved.

The notion that slavery for most of its history was seen as a "natural, necessary, and inevitable feature of the social world" is a complex one. For one thing, naturalness, necessity, and inevitability are not identical concepts. To provide a more detailed model of denaturalization, it is necessary to disambiguate what proponents of the mechanism have in mind when they claim that a certain practice or institution such as slavery was seen as a "natural, necessary, and inevitable feature of the social world." To disambiguate naturalness, I will propose three distinct interpretations of what could be meant when someone claims that a practice or institution is natural or necessary in order to defend the idea that it should not be changed. In doing so, I am offering a rational reconstruction of the different meanings that one could draw upon in defending the claim that some practice or institution is natural, in order to see which of these interpretations best fits existing discussions of denaturalization. Naturally, what people have in mind when they claim that a practice or institution is natural may be ill defined, confused, or inchoate, and so their claim may not fit neatly into any of the three categories described below. However, if such claims were to be better defined, made less confused, and clarified, then, I claim, they would fall into one of the following categories:

> *Impossibility*: To say that a practice is natural or necessary is to claim that it cannot be changed. This type of necessity can be understood easily in other domains. For instance, given our current understanding of the terms and current level of technology, it is impossible for a piglet to mature into a cow. If it is claimed that a practice or institution is natural or necessary in this sense of the term, then it follows from the principle

25 James, *The Black Jacobins*, ix; Drescher, *Abolition*, 174; and Popkin, *A Concise History of the Haitian Revolution*.

26 Drescher, *The Mighty Experiment*, 121, and *Abolition*, 260–64.

that ought implies can that one ought not to try to change that practice or institution.

*Costs*: To say that a practice is natural or necessary is to claim that attempts to change that practice will come with perhaps unbearably high costs. It could be that the practice or institution is functionally necessary to secure some desirable outcome or that there are not any viable alternatives for fulfilling this function, and thus attempting to change this practice or institution will lead to costs in the form of the desirable outcome not being achieved. It could also be the case that changing the practice will come with transition costs that are deemed too high.

*Natural Is Good*: To say that a practice is natural or necessary is to say that it is good. For instance, according to certain traditional Aristotelean views, finding out that the function of human sexual organs is to facilitate reproduction directly implies that the ethical purpose of human sexual activity is reproduction. With regard to slavery, David Brion Davis claims that "for the [ancient] Greeks (as for Saint Augustine and other early Christian theologians) physical bondage was part of the cosmic hierarchy, of the divine scheme for ordering and governing the forces of evil and rebellion."[27] More generally, cosmologies in hierarchical agricultural societies have often emphasized the divinely or cosmically ordained nature of hierarchical social institutions, such that challenging these institutions would be against the natural order of things and thus wrong.[28] These are examples of natural-is-good-type explanations for why practices or institutions are natural or necessary and thus should not be changed.

Which of these three interpretations do proponents of denaturalization have in mind? I argue that of these three interpretations, the costs interpretation is the best fit. For instance, when discussing the views that people have historically held about slavery, philosophers tend to emphasize the indispensable social role that slavery was thought to play in producing a social surplus. The idea is that people in slaveholding societies believed that, as a matter of functional necessity, without forced labor people would voluntarily work only enough to secure their own subsistence, and therefore there would be no social surplus. Without a social surplus, all forms of manufacturing that require investment,

27 Davis, *The Problem of Slavery in the Age of Revolution*, 42. But see Williams, *Shame and Necessity*, ch. 5 for a perspective that attributes this cosmological view mainly to Aristotle rather than to ancient Greek society at large.

28 Acemoglu and Johnson, *Power and Progress*, 121.

as well as the social roles of magistrates, clergy, educators, writers, artists, and scientists, could not be sustained. In combination, these claims amounted to the belief that slavery was necessary to sustain civilization.[29] Pleasants seems to hold this interpretation. He rejects the impossibility interpretation, most clearly in his discussion of the work of Michelle Moody-Adams. Moody-Adams attributes the impossibility interpretation to people who claim that perceptions of naturalness and necessity have upheld unjust social practices and institutions. She then argues that such claims must be bogus because it is not possible for competent language users to truly think that any of their social practices are necessary, because their ability to negate statements implies their ability to imagine social states in which any particular practice does not exist.[30] Pleasants (in my view rightly) responds that this is an implausibly strong interpretation of what it means to interpret some social practice as necessary, because it implies that any member of slaveholding society should have been willing to "give up slavery even if they believed that doing so would severely diminish the quality and viability of their society's way of life."[31] For Pleasants, claims about the necessity or naturalness of a practice amount to claims that there is no plausible alternative to the practice that is readily available and would not destabilize the social order and leave people "much worse off."[32] This is another example of what I have labeled the costs interpretation. As such, it seems that proponents of denaturalization claim that in the case of British abolitionism, denaturalization occurred because the alternative institution of wage labor enabled people (both those in positions of power and those in the broader public sphere who campaigned against slavery) to make their judgments about the costs of abandoning slavery more accurate: this cracked the veneer of naturalization.

For the rest of this paper, I will therefore adopt the costs interpretation as the understanding of what it means to claim that a practice is natural, necessary, or indispensable. However, before proceeding, a little more should be said about the natural-is-good interpretation. While I believe that costs and natural-is-good are conceptually distinct senses of naturalness, this does not mean that, on a psychological level, they are separate. It could well be that beliefs about an institution or practice being inevitable or very costly to abandon in

29 Anderson, "Social Movements, Experiments in Living, and Moral Progress," 16–17. Anderson is not claiming (and neither am I) that *if* this belief about the functionality of slavery was epistemically justified *then* the practice itself would be morally justified. Rather the claim is that this belief about the functionality of slavery had an effect on people's willingness to consider abandoning the practice.

30 Moody-Adams, *Fieldwork in Familiar Places*, 100.

31 Pleasants, "Moral Argument Is Not Enough," 169.

32 Pleasants, "Moral Argument Is Not Enough," 169.

a descriptive sense can foster beliefs about that institution or practice being morally good.[33] In that case, in order to fully understand historical instances of denaturalization, we need, in addition to a costs perspective, an account of how natural-is-good beliefs operate and how they can be overcome. Due to space constraints, I will focus only on an understanding of denaturalization that uses the costs interpretation, but this is not because I think that this is the only interpretation worthy of investigation.

As it stands so far, the idea that moral progress can be facilitated by people coming to have more accurate beliefs about the costs of abandoning their institutions is an intriguing one. However, this notion is currently vague. Exactly how should we understand these "costs"? How can we understand institutions being compared in terms of the benefits they provide and hence the costs of abandoning one to move to the other? And, given that the costs interpretation is a rational reconstruction of naturalness claims, is it psychologically realistic to think that people have something like these kinds of judgments about the costs of abandoning their institutions? In the following two sections, I will offer answers to these questions and, in doing so, develop a more detailed account of denaturalization.

## 2. Denaturalization as Improving Costs Judgments

Given the interpretation of naturalness settled on in the previous section, denaturalization occurs when an individual or group has some judgment, perhaps inchoate, about costs such that they believe getting rid of an institution will come with high costs, and then these judgments are rendered more accurate. This then facilitates a change away from that institution to a morally preferable one. Going forward, I will make use of the idea of a *costs judgment.* This is a judgment about the costs of moving from a status quo institution or practice to an alternative institution or practice. Naturally, much more needs to be said about how these costs of moving from one institution to another are to be understood. In this section, I will attempt to provide a more precise understanding of costs. I will argue that we can understand what costs judgments attempt to track using resources from the philosophy of conventionality.

33 See Jost, "A Quarter Century of System Justification Theory"; and Jost et al., "The Future of System Justification Theory." However, in section 4 below I will also explore the possibility of the opposite relationship obtaining, such that when an institution or social practice is denaturalized, this will incentivize people who benefit from that institution or social practice to produce moral justifications in its favor.

David Lewis analyzes conventions as equilibria in repeated *coordination games*.[34] Consider the following game in which the two players would like to coordinate their actions:

| | | Player 2 | |
|---|---|---|---|
| | | *A* | *B* |
| Player 1 | *A* | 1, 1 | 0, 0 |
| | *B* | 0, 0 | 1, 1 |

FIGURE 1 Simple coordination game

The game has two players (1 and 2) and two strategies (*A* and *B*) that yield certain payoffs. It is standard to interpret payoffs as representing preference rankings expressed in terms of *utility* in the rational choice sense of the term.[35]

In this game, the players are able to coordinate if they both choose the same strategy: if they either both play *A* or both play *B*. If the players coordinate, for example, by both playing *A*, then they have reached what Lewis refers to as a *proper coordination equilibrium*. In such a situation, neither player can improve their own payoff by unilaterally switching strategies, and neither player can improve the payoff for the other player by unilaterally switching strategies. Settling on *A*/*A* as a strategy is a convention because it is arbitrary: the players would have been just as well-off if they played *B*/*B* instead. However, if the game is played repeatedly, then once the *A*/*A* pattern emerges, it is a stable equilibrium because it is a proper coordination equilibrium: each player has a strong incentive to keep playing *A* because they cannot benefit themself or the other player by unilaterally switching to *B*.

Institutions can also be illuminated using this theoretical apparatus. Institutions can be modeled as sets of (often formalized) norms that, along with incentives and expectations, coordinate people's actions and thus stabilize patterns of behavior. Because of this stabilizing function, institutions can be understood as the (conventional) equilibria of repeated coordination games, as in figure 1.[36] Of course, a model of any actually existing institution would be vastly more complex than figure 1, involving many more players and many more possible outcomes.

34 Lewis, *Convention*.

35 Guala, *Understanding Institutions*, 21–22; and Gaus, *On Philosophy, Politics, and Economics*, ch. 2. See also Kogelmann, "What We Choose, What We Prefer," for a recent and sophisticated account of how to understand preference rankings.

36 Guala, *Understanding Institutions*, ch. 2.

Why think that looking at conventions can give us insight into how costs judgments can be modeled and that this insight might help us think more clearly about denaturalization? The institution of slavery was clearly conventional in the sense that it was a coordination equilibrium that could have been (and now is) otherwise. A pertinent question is how to understand the institution of slavery using the kind of model sketched above. If we model the institution of slavery as an equilibrium in a complicated coordination game, then are enslaved people counted as players in this game? Francesco Guala argues that slavery, seen as an equilibrium to a coordination game that includes enslaved people, can be seen as generally beneficial in the technical and circumscribed sense that the real alternative to being subject to the institution of slavery for many people throughout history has been being killed.[37] This claim seems to assume that enslavement is preferable to being killed according to the utility function of enslaved people. However, it is not clear that this claim is plausible. For one thing, Guala's characterization of alternatives may be inaccurate: in some contexts, the alternative to enslavement may not have been death or the risk of death but rather (the risk of) severe punishment. For another, even in cases where (the risk of) death was the alternative to enslavement, we have plenty of evidence that the demand for liberty from enslavement sometimes motivated enslaved people to take up arms against their enslavers in the face of fearsome odds of death, which suggests that the arrangement was not always beneficial even in Guala's circumscribed sense.[38]

When trying to use this understanding of institutions as the equilibria of repeated coordination games to understand the costs judgments of people who accepted slavery, I think it makes most sense to think of enslaved people as not being players in the game. The costs of abandoning slavery are thought to be costs for people who are not enslaved, and it is these perceived costs that affect the views and actions of people who directly benefit from or tolerate the institution of slavery. However, when considering whether denaturalization is always or generally morally progressive in section 4, this issue of who is included in the set of people whose costs judgments become more accurate will be very important.

I have now described a view of institutions according to which they can be modeled as the equilibria of repeated coordination games. These equilibria are conventional when they are arbitrary, and they are arbitrary when alternative coordination equilibria are possible. If we link this account of institutions to the description of denaturalization given in section 1, then we can say proponents

37 Guala, *Understanding Institutions*, 4–5.

38 James, *The Black Jacobins*; and Popkin, *A Concise History of the Haitian Revolution*.

of denaturalization hold that many past people had a false view of the institution of slavery, according to which it was, in some sense, not conventional: it was rather a "natural, necessary, and inevitable feature of the social world." Understanding more about what conventions are can help us understand the way in which past persons were mistaken, and this can help us understand how denaturalization, understood through the lens of the costs interpretation, might operate by correcting these mistakes.

Recent work by Mandy Simons, Kevin Zollman, and Cailin O'Connor provides this understanding by giving more insight into the notion of conventionality.[39] They suggest that the arbitrariness of a convention is not a binary matter. Instead, it can vary depending on three factors:

1. *Payoffs:* Some conventions have higher payoffs than others.[40]
2. *Stability:* Some conventions are more stable than other conventions in that they can tolerate a greater amount of deviance (people failing to play the conventional strategy) before the convention collapses to be replaced by another.
3. *Likelihood of Emergence:* Some conventions are more likely to emerge than others, either because there are only a small number of possible conventions or because some convention is more attractive to players due to higher payoffs, shared cultural norms, or cognitive biases.

To understand the first factor, consider the following game.[41]

| | | Player 2 | |
|---|---|---|---|
| | | *A* | *B* |
| Player 1 | *A* | 1, 1 | 0, 0 |
| | *B* | 0, 0 | *x*, *x* |

FIGURE 2 Coordination game in which *B*/*B* is the preferable equilibrium if $x > 1$.

Let us assume that $x = 100$. In this case, both *A*/*A* and *B*/*B* are proper coordination equilibria as defined above, and so they would both be candidates to be conventions on Lewis's account. However, if the players were to settle on

39 Simons and Zollman, "Natural Conventions and Indirect Speech Acts"; and O'Connor, *The Origins of Unfairness* and "Measuring Conventionality."

40 Another way of putting this is that some conventions are Pareto-superior to others. See Simons and Zollman, "Natural Conventions and Indirect Speech Acts," 7.

41 O'Connor, "Measuring Conventionality," 582.

the *B*/*B* equilibrium, then although their choice is arbitrary in that it *could* have been otherwise (*A*/*A* is also a proper coordination equilibrium), the explanation of why the players settled on *B*/*B* will likely involve an appeal to the much higher payoffs of *B*/*B*. Thus, while *B*/*B* is arbitrary in some sense, there is also a strong functional explanation available for why *B*/*B* might come to dominate as a strategy over *A*/*A*. Furthermore, if players who were playing *B*/*B* were asked to move from that equilibrium to *A*/*A*, then they could truly claim that that transition would come with very large costs because (again, assuming that $x = 100$) *A*/*A* has such low payoffs relative to *B*/*B*. Here we can see how a claim about payoffs can be related to a costs judgment.

Figure 2 can also help us understand the second factor, stability. If $x = 100$, then *A*/*A* will be a relatively unstable equilibrium. Why is this? Because if a population is playing *A*/*A*, then it will take only a relatively small percentage of the population defecting to playing *B*/*B* for the *A*/*A* equilibrium to collapse.[42] Regarding the third factor, there are several things that affect the likelihood of a convention emerging. For one thing, the likelihood of a given practice emerging depends on how many proper coordination equilibria exist with regard to that practice. For instance, imagine that figure 1 represents two possible conventions: driving on the left-hand side of the road and driving on the right-hand side of the road. Both conventions are proper coordination equilibria. Driving on the left-hand side of the road is arbitrary, but it is not *that* arbitrary because there is only one other proper coordination equilibrium: driving on the right-hand side. However, if we are dealing with a coordination game in which there are many different proper coordination equilibria (assuming, for now, that these equilibria have equivalent payoffs), then any given equilibrium will be more arbitrary simply because there are more possible alternative equilibria. Thus, we might say that the more proper coordination equilibria there are in a coordination game, the more arbitrary the emergence of any particular equilibrium is because there are more ways that this convention could have been otherwise.[43] The payoffs of a convention can also influence its likelihood of emerging, particularly due to the fact that a convention with higher payoffs is more likely to be adopted and more likely to spread from one social group to another.[44] Lastly, the likelihood of a convention emerging can be affected by perceptual, cognitive, or cultural biases that make a particular convention more salient for the relevant population.[45]

42 Simons and Zollman, "Natural Conventions and Indirect Speech Acts," 7–9; and O'Connor, "Measuring Conventionality," 584.

43 O'Connor, "Measuring Conventionality," 582.

44 Cohen, "Cultural Variation," 464; Henrich, *The WEIRDest People in the World*, 88–99; and O'Connor, "Measuring Conventionality," 584.

45 Guala, *Understanding Institutions*, 14–16.

We now have three factors that can influence the degree to which a practice or institution is conventional. How can this understanding of conventions inform our understanding of costs judgments? We can think of a costs judgment as the claim that changing an existing institution will result in drastically lower payoffs and/or that alternative institutions will be unstable and so unable to coordinate people's behavior in order to deliver acceptable payoffs. Thus, the first factor, payoffs, is directly relevant to the accuracy of a costs judgment: if a status quo institution provides the highest possible payoffs, and abandoning it will result in very low payoffs, then one can have an accurate costs judgment that abandoning that institution would come with heavy costs. This is a more abstract and precise way of articulating the kind of belief that Anderson, Jamieson, Hermann, and Pleasants attribute to people who thought that slavery was a natural, necessary, or indispensable institution: although, of course, in this case, the costs judgment was inaccurate. Stability is also relevant, because if an alternative institution is highly liable to defection and thus highly unstable, then this instability might result in significant costs when the institution collapses. This would make the alternative institution undesirable in terms of payoffs, relative to the status quo institution. The relevance of the third factor, the likelihood of emergence, is less clear. It seems relevant for costs judgments than an institution is likely to emerge because it has high payoffs, but this is just an indirect way of talking about the first factor. However, it does not seem directly relevant to assessing the costs of moving away from a given institution or practice that it is a convention that was highly likely to emerge due to the shared cognitive biases or cultural norms of the population that has that practice or institution. This would be relevant to a costs judgment only if these same cognitive biases or cultural norms mean that there would be costs involved in transitioning away from said institution. However, that the status quo institution is supported by shared cognitive biases or cultural norms may be very relevant for explaining why groups may be reticent to move away from the status quo, as will be explored further in section 3.

This model from the philosophy of conventionality gives us a clearer way of thinking about the features of practices and institutions that costs judgments attempt to track—namely, their payoffs and stability. If we have this understanding of costs judgments, then denaturalization would function by making them more accurate. Therefore, one important empirical assumption made by the account of denaturalization that I have developed is that people have judgments that, in some way, attempt to track the payoffs of their own institutions and social practices relative to alternatives. Fully developing an account of what these judgments are and how they attempt to track payoffs is too large a task to attempt in a paper of this length, although I will make a limited case for the

psychological realism of this account of denaturalization in section 3. For now, my point is that for the costs account of denaturalization sketched above to be a plausible causal mechanism of moral progress, we need a satisfactory account of what such payoff-tracking judgments are and how they work. Alternatively, we need to develop an alternative costs account of denaturalization to the one developed here that can explain what the relevant costs are and does not need an account of payoff-tracking judgments, or to develop an account of denaturalization that does not adopt the costs interpretation of naturalness and necessity but rather some other interpretation.

Assuming some psychological account of how people's costs judgments track the payoffs of institutions and social practices, how might costs judgments be made more accurate? According to proponents of denaturalization, exposure to existing alternative institutions can make costs judgments more accurate. Exposure to these alternative institutions can provide information about the payoffs and stability of alternatives, which can denaturalize the status quo institution by making it clear that abandoning this institution will not lead to unbearably high costs in terms of loss of payoffs. Once costs judgments are rendered more accurate, moral considerations can then play more of a role in motivating people to change their institutions. One implication is that the ability of people to improve the accuracy of their costs judgments is bounded by the actual alternative institutions that exist: without actual alternatives, one cannot assess the relative payoffs of alternatives to the status quo. On this account, people who tolerated or supported slavery before the emergence of widespread wage labor had an inaccurate costs judgment to the effect that a social surplus was not possible without slavery (which we now know is possible), but surveying existing alternative institutions at the time would not have provided the kind of information needed to update this costs judgment. Thus, this model of denaturalization implies that there are great benefits to engaging in institutional experimentation because such experiments in living are the only way to provide the evidence about payoffs and stability of alternative institutions that are vital to improve the accuracy of costs judgments and to potentially achieve denaturalization.[46]

46 On the value of institutional experimentation, see Anderson, "Social Movements, Experiments in Living, and Moral Progress"; Müller, "Large-Scale Social Experiments in Experimental Ethics"; and Robson, "The Rationality of Political Experimentation." Naturally, engaging in such experimentation may have diminishing returns, and the costs account of denaturalization says nothing about the opportunity costs of engaging in institutional experimentation. Nonetheless, the costs account does imply that there are strong *pro tanto* reasons to engage in institutional experimentation.

There is a complication worth noting here. Suppose that a given group forms the judgment, perhaps based on some small-scale institutional experiments, that moving to an alternative and more just institution will not be prohibitively costly for them, and this judgment makes moral criticism of the status quo institution more effective and facilitates a transition to a new institution. However, it then transpires that this costs judgment was wrong. Moving to this new institution, while it is *ex hypothesi* more just, has much lower payoffs for them than the status quo institution, such that the institutional change is perceived to be prohibitively costly. In this case, what cracked the veneer of naturalization of the status quo institution was not that the group in question came to have more accurate beliefs about the costs of moving to an alternative institution but rather that they *believed* that moving to the alternative institution would not have prohibitive costs.[47] Anderson points out that in the case of British abolitionism, a group of British elites extrapolated their judgments about the payoffs of abolishing slavery based on small-scale experiments in abolition (as described in section 1), but for at least some of these people, their expectations of increased productivity in the lucrative British sugar colonies of the Caribbean following abolition (better payoffs from the new institution as compared to the old) were disappointed.[48] In other words, their belief about improved payoffs from moving to an alternative institution was false, but this belief still facilitated a transition to a more just institution. So do more accurate costs judgments really matter for facilitating institutional change, or is what matters simply that people who would otherwise resist those changes come to believe that those changes will not be prohibitively costly for them, even if they are wrong?

I believe that more accurate costs judgments are in fact important if durable institutional change is to be obtained. If people have mistakenly optimistic judgments about the costs of moving to alternative institutions as described above, then while this may facilitate institutional change, it is also likely to lead to backlash once it becomes clear that the new institution has prohibitively high costs. I submit that institutional change is likely to be more durable if people's projections of the costs of moving to alternative institutions are at least relatively accurate, so that it is true that the more just institutions are not prohibitively unstable and do not deliver unacceptably low payoffs. Returning to the example of British abolitionism, this was by and large the case. Despite the mistaken beliefs described by Anderson of some British elites regarding the relative productivity of wage labor versus that of slave labor, Pleasants makes clear that the "abandonment of slavery for the newly emerging paradigm of freely

47 I thank an anonymous reviewer for drawing my attention to this kind of case.

48 Anderson, "Social Movements, Experiments in Living, and Moral Progress," 18–20.

contracted wage labour served the medium- to long-term economic interest of the liberators spectacularly well."[49] In the medium to long term, wage labor as an alternative institution to slavery did not deliver unacceptably low payoffs, and the fact that this *was* the case (as opposed to people merely projecting, wrongly, that it *would* be the case) can reasonably be thought to have played a role in ensuring that the morally progressive transition from slavery to wage labor has been sustained.

When forming costs judgments about moving from a status quo institution to a more just institution based on small-scale institutional experiments, we must recognize that our costs judgments are always going to be projections, and we will not know whether our costs judgments have truly become more accurate except in hindsight. However, if I am correct about the importance of more accurate costs judgments, then this implies that great attention should be given to the potential pitfalls of extrapolating incorrect predictions from small-scale experiments with alternative institutions because if our costs judgments only appear to have become more accurate rather than really becoming so, then this could facilitate unstable moral progress and dangerous backlashes.

I have now explicated an account of denaturalization, the costs account, that is more detailed than the descriptions of denaturalization thus far offered in the literature. My account is explicit about the interpretation of naturalness being used, shows how this kind of naturalness can be understood using resources from the philosophy of conventionality, and shows how people can be mistaken about the naturalness of their institutions in degrees. However, in Popperian fashion, making the hypothesis that denaturalization is a causal mechanism of moral progress more detailed and specific does not necessarily make it more convincing; instead, it brings into sharp relief the various points of criticism that can be leveled against the account. I see this as an entirely good thing, if one's aim is to advance our knowledge about this proposed mechanism of moral progress. In the following section, I will add more detail to the account by making a brief case for its psychological realism.

## 3. The Psychological Realism of Denaturalization

While the costs interpretation of denaturalization is a rational reconstruction, it is nonetheless the case that denaturalization is meant to at least partially explain real processes of historical change. For this to be plausible, it must be the case that the costs interpretation is rooted in some real psychological mechanisms that explain people's behavior. What needs to be established in order to believe

49 Pleasants, "The Structure of Moral Revolutions," 591.

that the account of denaturalization offered above is psychologically realistic? We would need to establish that people have psychological states that are similar to what I have been calling costs judgments—judgments that attempt to track the payoffs and stability of their institutions and practices relative to available alternatives.[50] Further, we should have evidence that people's costs judgments can become more accurate through being exposed to alternative institutions and practices: this would be evidence that experiments in living can provide correctives to inaccurate costs judgments, thus denaturalizing status quo institutions.

In this section, I will provide some evidence for the psychological realism of my account of denaturalization, with the proviso that more evidence would need to be provided to truly vindicate the account. Nonetheless, this section provides a sampling of the kind of evidence needed to support an account of denaturalization like the one outlined in section 2 or any similar account that takes the costs interpretation of naturalness described in section 1.

Firstly, do humans actually keep track of the payoffs and stability of their institutions relative to alternatives? Evidence from anthropology and cultural evolutionary theory suggests that they do. One source of evidence is research on *subjective selection*. Subjective selection refers to the selective retention of beliefs, practices, and other cultural variants that people subjectively evaluate as being useful, especially for fulfilling their goals.[51] In addition to explaining how people selectively retain or reject things like hunting practices and tools, subjective selection also affects the selective retention of rules and norms that are perceived to satisfy the interests of those who are in positions to build, maintain, and enforce rules and norms.[52] As a mechanism of cultural change, subjective selection requires that people have psychological states that track the subjective costs and benefits of different beliefs and practices. These psychological states are similar to those that I have described as costs judgments.

Another source of evidence comes from research on intergroup competition. Joseph Henrich describes how cultural evolution can give rise to packages of prosocial norms and institutions through a process of intergroup competition.[53] There are numerous ways in which competition between groups with

50 That people have these kinds of psychological states is an important presupposition of the costs account of denaturalization. If, instead, people typically do not make such assessments of status quo institutions, then this would count against the costs interpretation.

51 Singh, "Subjective Selection and the Evolution of Complex Culture," 266.

52 Singh, "Subjective Selection and the Evolution of Complex Culture," 267, 272–73; and Singh et al., "Self-Interest and the Design of Rules."

53 By 'prosocial', Henrich means norms and institutions that lead to success in intergroup competition, for instance by fostering cooperation or internal harmony within the in-group. See Henrich, *The Secret of Our Success*, 169.

different norms and institutions can lead to the spread of more prosocial norms and institutions, but two in particular are relevant for the purposes of this paper: *differential migration* and *prestige-based group transmission*.

Differential migration describes the process in which individuals preferentially migrate to more successful groups whose norms and institutions create "greater internal harmony, cooperation, and economic production."[54] Of course, greater internal harmony, higher levels of cooperation, and greater economic production are all things that contribute to higher payoffs and greater stability in the senses described in the previous section.[55] This suggests that people who are migrating preferentially to more successful groups have judgments that, at least to a large extent, track the payoffs and stability of the institutions and practices of the group that they migrate to relative to the institutions and practices of their original group. These judgments appear to approximate costs judgments.

Prestige-based group transmission occurs when individuals in one group preferentially attend to and copy the social norms of other, more successful, groups.[56] Where the individuals in the copying group also have the ability to legislate norm and institution change for their entire group, this can also result in an entire group adopting the norms and institutions of a more successful group. Henrich offers the example of a community in New Guinea called Ilahita who in the late nineteenth century copied a package of rituals, religious beliefs, norms, and institutions (collectively called the *Tambaran*) from a militarily successful group called the *Abelam*, whose expansion was a potential threat to Ilahita. The Tambaran was already being adhered to by the Abelam, and it was thought by Ilahita's elders that the Tambaran was the source of the Abelam's success. By copying the Tambaran and making some felicitous errors in how they copied it, Ilahita ended up not only matching but surpassing the military might, level of cooperation, and scale of the Abelam.[57] Prestige-based group transmission suggests that people within groups have judgments about the relative payoffs (often in terms of military might or level of cooperation) of their institutions and practices and the institutions and practices of other groups, and where the institutions or practices of other groups are superior, people are sometimes motivated to copy them.[58] These judgments also appear to approximate costs judgments.

54 Henrich, *The Secret of Our Success*, 168.

55 Heath, "The Benefits of Cooperation."

56 Henrich, *The Secret of Our Success*, 168.

57 Henrich, *The WEIRDest People in the World*, 88–99.

58 One crucial caveat about prestige-based group transmission is that the link between the practices and institutions of other groups and the desirable higher payoffs of these practices and institutions is often causally opaque: it is not clear which practices or institutions

Additionally, we need to explain how costs judgments can be made more accurate through exposure to alternative institutions. In part, this explanation is provided by the above account of forms of intergroup competition in which people acquire information about the payoffs of alternative institutions. However, denaturalization is meant to work by correcting inaccurate costs judgments. What factors could make costs judgments inaccurate, such that denaturalization can then act to make them more accurate? Firstly, people could simply lack knowledge about other possible institutions that have equivalent or higher payoffs than their status quo institutions. Secondly, the cultural evolutionary framework referred to above may support the idea the people have something like costs judgments, but it also suggests that humans have a norm psychology that makes social norms and institutions difficult to change because people are often intrinsically motivated to follow the norms that they grew up with and to punish norm violations. Punishment can then render systems of norms stable against shocks, including deliberate attempts to change such systems.[59] To the extent that people's intrinsic motivation to follow their status quo norms and their motivation to punish norm violations can bias their perception of the costs of changing their status quo norms, practices, or institutions, these factors could contribute to explaining why costs judgments can be inaccurate.

Thirdly, people could underestimate the payoffs of moving to an alternative practice or institution and thus overestimate the costs of moving from the status quo to the alternative. This possibility is suggested by the phenomenon of *loss aversion*, in which the risks of loss associated with changing away from the status quo can weigh much more heavily in people's minds than the prospective gains associated with change—a particularly important error when it comes to making accurate costs judgments.[60] Loss aversion has recently been challenged on a number of grounds: that much of the evidence for loss aversion has been overinterpreted because there are other interpretations of these results that do not support the existence of loss aversion, and that whether or not losses are weighed more heavily depends on the context of choice.[61] But

are causally responsible for the perceived success. As a result, when people choose to copy the practices or institutions of other groups, they tend to copy quite indiscriminately, adopting many such practices and institutions rather than adopting only the ones that contribute to the higher payoffs in a targeted way (Henrich, *The WEIRDest People in the World*, 97).

59 Kelly and Davis, "Social Norms and Human Normative Psychology," 63–64; Henrich, *The Secret of Our Success*, ch. 9; and Boyd and Richerson, "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups."

60 For classic descriptions of loss aversion, see Kahneman and Tversky, "Choices, Values, and Frames"; and Kahneman et al., "Anomalies."

61 Gal and Rucker, "The Loss of Loss Aversion"; and Yechiam, "Acceptable Losses."

more recently, high-powered studies have demonstrated that loss aversion is a robust phenomenon, even when dealing with small losses, but also that loss aversion has moderators: there are some features of decision makers that can attenuate loss aversion.[62] More educated decision makers are less prone to loss aversion than less educated ones, older decision makers are more prone to loss aversion than younger ones, and people with more experience and knowledge about the decision domain in question are less prone to loss aversion than those with less experience. This last moderator in particular suggests that experience with relevant alternatives can aid in making costs judgments more accurate by mitigating loss aversion, which bolsters the case that institutional experimentation can contribute to denaturalization.

The evidence presented in this section makes a preliminary case for the psychological realism of my account of denaturalization by arguing that people have psychological states that approximate costs judgments; that there are psychological factors, including how human norm psychology works and our vulnerability to loss aversion, that can explain why costs judgments can be inaccurate; and that exposure to alternative institutions can make costs judgments more accurate. Given the brevity of this presentation of evidence, we of course cannot say conclusively whether the account is psychologically realistic. However, this section nonetheless gives an indication of the kind of evidence that would be needed to demonstrate that an account of denaturalization (especially one based on some version of the costs interpretation) is realistic. Future accounts of denaturalization should try to provide similar and ideally more advanced evidence for their psychological realism.

## 4. Denaturalization and Moral Progress

So far, I have analyzed denaturalization as it has been proposed in the literature; argued that denaturalization works by making costs judgments more accurate; provided a model of how we can understand what costs judgments aim to track; and provided evidence that my account of denaturalization possesses a degree of psychological realism. Taken together, this gives us an account of denaturalization that is more detailed and specific in its claims than previous discussions of denaturalization in the literature. I hope that this account can be critically assessed and improved upon in future philosophical work.

In this last section, I will assume that the costs account of denaturalization is correct in order to situate denaturalization as a cause of moral progress within

62 Ruggeri et al., "Replicating Patterns of Prospect Theory for Decision Under Risk"; and Mrkva et al., "Moderating Loss Aversion." In Ruggeri et al., $n = 4{,}098$ participants from nineteen countries; and in Mrvka et al., $n = 17{,}720$ across five unique samples.

the broader moral progress literature and attempt to answer a key question: Will denaturalization always or even generally lead to moral progress? After all, it could instead be a mechanism of moral change with a random moral valence or, worse, be generally biased in favor of morally regressive social change.

Before getting started, let us briefly consider the question of what it means for something to be morally progressive. Firstly, I will assume that certain cases are canonical examples of moral progress that are beyond reasonable doubt—including the abolition of slavery, gains in gender equality, and increasing recognition of the moral acceptability of same-sex relationships.[63] Secondly, I will assume that all human beings have equal moral status. Given this moral standard, social changes that result in this belief being more widely held and, correspondingly, result in people being treated equally regardless of group membership will count as moral progress.[64]

If denaturalization was a contributing cause of the British abolition of slavery, then it is hard to doubt that it was morally progressive in that specific case. However, in general, whether denaturalization will lead to moral progress depends on a number of factors. Firstly, recall that denaturalization works by making costs judgments more accurate so that a switch to an alternative institution is no longer (falsely) thought to have unacceptably high costs. With this false belief removed, moral criticism of the status quo institution can then be more effective in mobilizing change. According to this story, denaturalization alone is not sufficient for moral progress. Justified moral beliefs or values are also necessary to motivate the change away from the status quo institution and towards the morally preferable one. Thus, denaturalization can facilitate moral progress when inaccurate costs judgments that are contributing to the inefficacy of justified moral criticism are removed, but this justified moral criticism is still necessary for denaturalization to facilitate progress.

Secondly, assuming that people have justified moral beliefs or values, whether denaturalization can facilitate progress depends on the actual payoffs of alternative institutions relative to the status quo. If we imagine that in fact there were no alternatives to the institution of slavery for producing a social surplus, then if people who benefitted from or tolerated slavery came to have more accurate costs judgments, this would not facilitate progress. Rather, it

63 Buchanan and Powell, *The Evolution of Moral Progress*, 47–48, 241; Buchanan, *Our Moral Fate*, xiii; Kitcher, *Moral Progress*, 13; and Kumar and Campbell, *A Better Ape*, 181.

64 Buchanan and Powell, *The Evolution of Moral Progress*, 11–18. Questions can certainly be asked about how the standards for moral progress are justified. However, for the purposes of exploring how the denaturalization mechanism relates to the overall philosophy of moral progress, I will rely on these moral standards, which are already widely accepted in the moral progress literature.

would entrench the belief that slavery could not be abandoned without high costs. In such a case, it would better facilitate moral progress if such people came to have even more inaccurate costs judgments so that they falsely believed that alternative institutions had comparable or higher payoffs to their status quo slave institutions (though, as mentioned in section 2, such moral progress based on inaccurate costs judgments would likely be unstable). Victor Kumar and Richmond Campbell argue, paraphrasing Stephen Colbert, that "reality has an inherent progressive bias" such that when people come to have more accurate beliefs about the world around them, they tend to modify their moral norms and values in the direction of inclusion, equality, and progress.[65] For denaturalization to be reliably progressive, it must be the case that this is by and large true, so that coming to have more accurate costs judgments about the relative payoffs of unjust status quo institutions and relatively more just alternative institutions has the effect of making the status quo seem less natural, inevitable, and necessary rather than entrenching this impression. Whether this *is* largely true is a difficult question to answer: it seems like something that rather needs to be considered on a case-by-case basis. Nonetheless, it seems to be the case that whether denaturalization can facilitate progress is largely hostage to whether the facts are such that there really are more just and roughly equivalent payoff institutions. These facts in turn are influenced by factors such as:

- Which institutions happen to be available as actual alternatives, which may largely be a matter of historical happenstance.
- What the other institutions and social norms of the people who are making costs judgments are. This is important because the payoffs of any given institution or practice depends to some extent on the culture (which includes the other institutions, practices, beliefs, and social norms) of the people who will be adopting them. Because of this, there is a certain path dependency whereby some institutions that might be highly effective for one group may be much less effective for another.[66]
- What kind of technologies are available, as technologies can also alter the payoffs of different social norms and institutions.[67]

These factors, at least, are important for working out whether, given justified moral values and beliefs, denaturalization can facilitate moral progress.

Thirdly, let us return to a point briefly made in section 2 about who is in the group from whose perspective costs judgments are being made. When we

65 Kumar and Campbell, *A Better Ape*, 195.

66 Henrich, *The WEIRDest People in the World*, 98, 476–78.

67 Hopster et al., "Pistols, Pills, Pork and Ploughs," 21–22.

consider the story of British abolitionism endorsed by proponents of denaturalization, the people whose costs judgments mattered were the antislavery campaigners and the political elites in Britain, because these were the people whose beliefs were causally efficacious in legislating the end of legal slavery. In this situation, it is fortunate that *that* rather limited group updated their costs judgments to believe that they would not experience unbearably low payoffs if they switched from their unjust status quo institution. But it is easy to imagine cases in which switching from an unjust status quo institution to a more just alternative institution will lead to higher or equivalent payoffs for the majority of people affected by the status quo institution but will lower the payoffs of the group who have decision-making power to effect that switch. In this case, updating the costs judgments of that group would not facilitate moral progress because updated costs judgments, even if they showed that an unjust institution could be abandoned without significantly lowering payoffs for the majority of people affected by the institution, would not be likely to result in any institutional change. Thus, it seems that denaturalization is more likely to facilitate moral progress the more inclusive the group that gains more accurate costs judgments is and the more inclusive the decision-making procedures to secure institutional change are. So, broadly speaking, we should expect denaturalization to work better in a context of inclusive morality, where many people's interests and moral status are equally respected, and inclusive institutions, in which many people whose interests are affected by those institutions have decision-making power within them or, at the limit, have an ability to influence those with decision-making power (as was the case with petitioners during the campaigns for abolition in Britain).[68]

However, I think there is also an interesting feedback loop between the inclusivity of social norms and institutions and the effectiveness of denaturalization as a mechanism of moral progress. British abolitionism led to an expansion of the moral circle and a gain in moral inclusivity through the recognition of a basic level of moral status and securing a basic level of legal status for formerly enslaved persons, but this gain in inclusivity was driven by a non-inclusive group that was numerically dominated by non-enslaved people.[69] If

68 On the importance of equality of moral status and respect, see Buchanan and Powell, *The Evolution of Moral Progress*, 62–64; Buchanan, *Our Moral Fate*, 23–24; and Kumar and Campbell, *A Better Ape*, 184–86. On inclusive institutions, see Acemoglu and Robinson, *Why Nations Fail*, 79–83. And on the position of petitioners in the British abolition movement, see Anderson, "Social Movements, Experiments in Living, and Moral Progress," 10–15.

69 Kumar and Campbell, *A Better Ape*, 203–7; and Buchanan and Powell, *The Evolution of Moral Progress*, 57, 212–14.

denaturalization can lead to gains in inclusivity, and gains in inclusivity can then increase the likelihood that further denaturalization will lead to moral progress, then denaturalization as a mechanism of moral progress and gains in moral inclusivity as a form of moral progress may form a positive feedback loop.

Fourthly, a less morally welcome feedback loop is that successful instances of denaturalization may give rise to ideologically motivated justifications for the moral rightness of an unjust practice. Imagine that within a slaveholding society, the group of people who are either slaveholders or who tolerate slavery come to see that moving to an alternative and more just institution will not result in prohibitively high costs—for example, it will still be possible to produce a social surplus without slavery. This denaturalization will then make moral criticism of slavery more effective. Even if this is the case, it is still going to be the case that some within the group will lose substantial benefits that they currently enjoy if slavery is abolished. Supposing that slavery has been denaturalized such that it is no longer plausible that it is a natural and necessary institution (according to the costs understanding of this claim), these people will no longer be able to make uncontested claims about the naturalness of slavery as an institution without alternatives. But this does not mean that this group will no longer have an interest in slavery continuing. Rather, it means that they need to produce justifications in favor of maintaining slavery. Indeed, as described in section 1, some historians have argued that explicit moral justifications for slavery emerged only late in the history of the institution—around the time that slavery was being denaturalized by the emergence of wage labor as an alternative institution. It is plausible that many instances of denaturalization will leave some members of the group that undergoes that denaturalization with strong interests in maintaining the status quo institution and thus with strong interests in producing moral justifications for the denaturalized status quo institution. These moral justifications will be ideological in the sense that they are epistemically distorted, in this case by the self-interest of the members of the group producing them.[70] Such ideologically distorted purported moral justifications for unjust institutions may commonly emerge in the wake of morally progressive denaturalization.[71]

To sum up, it seems that denaturalization is not a mechanism that is guaranteed to facilitate moral progress. Whether denaturalization will lead to moral progress depends on the factors enumerated above: whether there are justified moral beliefs and values that will correctly identify unjust status quo

70 Barrett, "Ideology Critique and Game Theory," 714n1.

71 Thanks to an anonymous reviewer for prompting me to discuss this phenomenon in greater detail.

institutions and push for their removal after they are denaturalized; whether it is in fact the case that there are more just alternative institutions with equivalent or higher payoffs available; how inclusive the group whose costs judgments are rendered more accurate is and how many members of that group have access to decision-making power to change the unjust status quo; and whether and to what extent particular social groups are able to produce successful ideological moral justifications of the unjust status quo in the face of denaturalization.

## 5. CONCLUSION

Moral progress, to the extent that it occurs, is likely to evade simple monocausal explanations.[72] In that spirit, this paper can be taken as an investigation into one of the many mechanisms that have been proposed to explain past instances of moral progress and that could potentially lead to future moral progress.

I have articulated a more detailed understanding of denaturalization than has thus far been offered in the literature, so that the mechanism can be critically assessed on empirical and philosophical grounds. I have argued that denaturalization works by improving our costs judgments and that these judgments are accurate to the extent that they track the relative payoffs and stability of different institutions. I have also provided evidence for the psychological realism of this account of denaturalization, both to bolster the case for my account and to show what kind of empirical evidence would be required to make the case that denaturalization is psychologically realistic. I hope that this developed account can be critically assessed by other philosophers interested in the mechanisms of moral change and moral progress and that it can encourage the development of further accounts of denaturalization—understood as improving costs judgments, understood as a mechanism that corrects false beliefs that fit into the natural-is-good interpretation outlined in section 1, or understood in some other way. Finally, with a more detailed account of denaturalization in hand, I have investigated its potential to facilitate moral progress and laid out the factors that affect whether denaturalization is progressive after all.[73]

*Utrecht University*
*c.t.blunden@uu.nl*

72 Eriksen, "The Dynamics of Moral Revolutions." For an account of some of the difficulties that are faced by accounts of what causes moral progress, see also Rehren and Blunden, "Let's Not Get Ahead of Ourselves."

73 Many thanks to Joel Anderson, Joseph Heath, Benedict Lane, Paul Rehren, and Hanno Sauer for discussing these ideas with me and providing criticism and feedback. I would like to extend special thanks to Chiara Cecconi for inviting me to present a draft of this paper

## REFERENCES

Acemoglu, Daron, and Simon Johnson. *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity*. Basic Books, 2023.

Acemoglu, Daron, and James A. Robinson. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Currency, 2012.

Anderson, Elizabeth. "Social Movements, Experiments in Living, and Moral Progress: Case Studies from Britain's Abolition of Slavery." Lindley Lecture, University of Kansas, 2014.

Barrett, Jacob. "Ideology Critique and Game Theory." *Canadian Journal of Philosophy* 52, no. 7 (2022): 714–28.

Boyd, Robert, and Peter J. Richerson. "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups." *Ethology and Sociobiology* 13, no. 3 (1992): 171–95.

Brown, Christopher Leslie. *Moral Capital: Foundations of British Abolitionism*. University of North Carolina Press, 2006.

Buchanan, Allen. *Our Moral Fate: Evolution and the Escape from Tribalism*. Massachusetts Institute of Technology Press, 2020.

Buchanan, Allen, and Rachell Powell. *The Evolution of Moral Progress: A Biocultural Theory*. Oxford University Press, 2018.

Cambiano, Giuseppe. "Aristotle and the Anonymous Opponents of Slavery." *Slavery and Abolition* 8, no. 1 (1987): 22–41.

Cohen, Dov. "Cultural Variation: Considerations and Implications." *Psychological Bulletin* 127, no. 4 (2001): 451–71.

Davis, David Brion. *The Problem of Slavery in the Age of Revolution, 1770–1823*. 2nd ed. Oxford University Press, 1999.

Drescher, Seymour. *Abolition: A History of Slavery and Antislavery*. Cambridge University Press, 2009.

———. *The Mighty Experiment: Free Labor versus Slavery in British Emancipation*. Oxford University Press, 2002.

Eriksen, Cecilie. "The Dynamics of Moral Revolutions: Prelude to Future Investigations and Interventions." *Ethical Theory and Moral Practice* 22, no. 3 (2019): 779–92.

---

to the History of Philosophy Colloquium at Utrecht University and to the participants at this colloquium for their engagement and their robust and incisive criticism, which were useful for the further development of the ideas in this paper. Thanks also to audiences at the IV GECOPOL Geneva Graduate Conference in Political Philosophy and the SOPHIA 2023 Conference. Thanks to the European Research Council (grant number 851043) for funding my research. Lastly, I would like to thank my two anonymous reviewers, the editorial team, and the copyeditor at the *Journal of Ethics and Social Philosophy*.

Gal, David, and Derek D. Rucker. "The Loss of Loss Aversion: Will It Loom Larger Than Its Gain?" *Journal of Consumer Psychology* 28, no. 3 (2018): 497–516.

Gaus, Gerald F. *On Philosophy, Politics, and Economics*. Thomson Wadsworth, 2007.

Guala, Francesco. *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton University Press, 2016.

Haskell, Thomas L. "Capitalism and the Origins of the Humanitarian Sensibility: Part 1." *American Historical Review* 90, no. 2 (1985): 339–61.

———. "Convention and Hegemonic Interest in the Debate over Antislavery: A Reply to Davis and Ashworth." *American Historical Review* 92, no. 4 (1987): 829–78.

Heath, Joseph. "The Benefits of Cooperation." *Philosophy and Public Affairs* 34, no. 4 (2006): 313–51.

Henrich, Joseph. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press, 2016.

———. *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. Farrar, Straus and Giroux, 2020.

Hermann, Julia. "The Dynamics of Moral Progress." *Ratio* 32, no. 4 (2019): 300–11.

Holslag, Jonathan. *A Political History of the World: Three Thousand Years of War and Peace*. Pelican, 2018.

Hopster, Jeroen, Chirag Arora, Charlie Blunden, Cecilie Eriksen, Lily Frank, Julia Hermann, Michael Klenk, Elizabeth O'Neill, and Stephen Steinert. "Pistols, Pills, Pork and Ploughs: The Structure of Technomoral Revolutions." *Inquiry* (2022): 1–33.

Huemer, Michael. "A Liberal Realist Answer to Debunking Skeptics: The Empirical Case for Realism." *Philosophical Studies* 173, no. 7 (2016): 1983–2010.

Jaeggi, Rahel. *Critique of Forms of Life*. Translated by Ciaran P. Cronin. Belknap Press, 2018.

James, C. L. R. *The Black Jacobins: Toussaint L'Ouverture and the San Domingo Revolution*. 2nd rev. ed. Vintage Books, 1989.

Jamieson, Dale. "Slavery, Carbon, and Moral Progress." *Ethical Theory and Moral Practice* 20, no. 1 (2017): 169–83.

Jost, John T. "A Quarter Century of System Justification Theory: Questions, Answers, Criticisms, and Societal Applications." *British Journal of Social Psychology* 58, no. 2 (2019): 263–314.

Jost, John T., Vivienne Badaan, Shahrzad Goudarzi, Mark Hoffarth, and Mao Mogami. "The Future of System Justification Theory." *British Journal of*

*Social Psychology* 58, no. 2 (2019): 382–92.
Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias." *Journal of Economic Perspectives* 5, no. 1 (1991): 193–206.
Kahneman, Daniel, and Amos Tversky. "Choices, Values, and Frames." *American Psychologist* 39, no. 4 (1984): 341–50.
Kelly, Daniel, and Taylor Davis. "Social Norms and Human Normative Psychology." *Social Philosophy and Policy* 35, no. 1 (2018): 54–76.
Kitcher, Philip. *Moral Progress*. Edited by Jan-Christoph Heilinger. Oxford University Press, 2021.
Kogelmann, Brian. "What We Choose, What We Prefer." *Synthese* 195, no. 7 (2018): 3221–40.
Kumar, Victor, and Richmond Campbell. *A Better Ape: The Evolution of the Moral Mind and How It Made Us Human*. Oxford University Press, 2022.
Lewis, David. *Convention: A Philosophical Study*. Harvard University Press, 1969.
Moody-Adams, Michele M. *Fieldwork in Familiar Places: Morality, Culture, and Philosophy*. Harvard University Press, 1997.
Mrkva, Kellen, Eric J. Johnson, Simon Gächter, and Andreas Herrmann. "Moderating Loss Aversion: Loss Aversion Has Moderators, but Reports of Its Death Are Greatly Exaggerated." *Journal of Consumer Psychology* 30, no. 3 (2020): 407–28.
Müller, Julian F. "Large-Scale Social Experiments in Experimental Ethics." In *Experimental Ethics: Towards an Empirical Moral Philosophy*, edited by Christoph Luetge, Hannes Rusch, and Matthias Uhl. Palgrave Macmillan, 2014.
O'Connor, Cailin. "Measuring Conventionality." *Australasian Journal of Philosophy* 99, no. 3 (2021): 579–96.
———. *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford University Press, 2019.
Pleasants, Nigel. "Moral Argument Is Not Enough: The Persistence of Slavery and the Emergence of Abolition." *Philosophical Topics* 38, no. 1 (2010): 159–80.
———. "The Structure of Moral Revolutions." *Social Theory and Practice* 44, no. 4 (2018): 567–92.
Popkin, Jeremy D. *A Concise History of the Haitian Revolution*. Blackwell, 2012.
Rehren, Paul, and Charlie Blunden. "Let's Not Get Ahead of Ourselves: We Have No Idea if Moral Reasoning Causes Moral Progress." *Philosophical Explorations* 27, no. 3 (2024): 351–69.
Robson, Gregory. "The Rationality of Political Experimentation." *Politics,*

*Philosophy and Economics* 20, no. 1 (2021): 67–98.
Ruggeri, Kai, Sonia Alí, Mari Louise Berge, Giulia Bertoldo, Ludvig D. Bjørndal, Anna Cortijos-Bernabeu, Clair Davison, et al. "Replicating Patterns of Prospect Theory for Decision Under Risk." *Nature Human Behaviour* 4, no. 6 (2020): 622–33.
Sauer, Hanno, Charlie Blunden, Cecilie Eriksen, and Paul Rehren. "Moral Progress: Recent Developments." *Philosophy Compass* 16, no. 10 (2021): e12769.
Simons, Mandy, and Kevin J. S. Zollman. "Natural Conventions and Indirect Speech Acts." *Philosophers Imprint* 19, no. 9 (2019): 1–26.
Singh, Manvir. "Subjective Selection and the Evolution of Complex Culture." *Evolutionary Anthropology* 31, no. 6 (2022): 266–80.
Singh, Manvir, Richard Wrangham, and Luke Glowacki. "Self-Interest and the Design of Rules." *Human Nature* 28, no. 4 (2017): 457–80.
Williams, Bernard. *Shame and Necessity*. University of California Press, 1993.
Yechiam, Eldad. "Acceptable Losses: The Debatable Origins of Loss Aversion." *Psychological Research* 83, no. 7 (2019): 1327–39.

*Journal of Ethics and Social Philosophy*
VOL. 29, NO. 2 · JANUARY 2025
*https://doi.org/10.26556/jesp.v29i2.3012*


# GRATITUDE FOR WHAT WE ARE OWED

*Aaron Eli Segal*

GRATITUDE occupies a central place in our moral landscape. We tend to feel gratitude when others benefit us out of good will, and we tend to express gratitude to others out of our recognition and appreciation of such good will. Strawson claims in "Freedom and Resentment" that gratitude, in playing this role, stands opposed to the reactive attitude of resentment, which we feel in response to displays of ill will.[1] But many hold that gratitude and resentment stand opposed to one another not just in relation to good and ill will but also in their relation to the demands of morality. Concerning resentment, many hold that *A* is warranted in resenting *B* only if *B* wrongs *A*, i.e., if *B* treats *A* in a way that *B* owes it to *A* not to treat them.[2] And further, many philosophers hold that gratitude likewise has an important connection to what we owe to each other: *A* never owes *B* gratitude for *B*'s treating *A* in a way that *B* owes it to *A* to treat them.[3] I will call this latter claim the *Orthodox Thesis*. These two claims about the relationship between gratitude, resentment, and what we owe to each other jointly characterize a conception of the role of good and ill will in interpersonal morality: ill will is displayed in someone's failing to live up to the demands of morality in their treatment of us, while good will is displayed in someone's going above and beyond the demands of morality in their treatment of us.

In the first part of this paper, I argue that the Orthodox Thesis is false—or at least that its scope must be restricted in an important way if it is to be plausibly maintained. That is, I argue that we sometimes owe others gratitude for treating us in ways that we are morally owed or, equivalently, for treating us in ways that we have a claim to.[4] I begin by presenting a range of cases that, I

1 Strawson, "Freedom and Resentment."

2 See, for instance, Wallace, *Responsibility and the Moral Sentiments*; and Darwall, *The Second-Person Standpoint*.

3 See Camenisch, "Gift and Gratitude in Ethics"; Lyons, "The Odd Debt of Gratitude"; Weiss, "The Moral and Social Dimensions of Gratitude"; Feinberg, "The Nature and Value of Rights"; and most recently, Macnamara, "Gratitude, Rights, and Benefit."

4 Some philosophers put this view in terms of owing others gratitude for their respecting our "rights." See, e.g., Macnamara, "Gratitude, Rights, and Benefit." I will avoid the term 'right' due to complications concerning the relation between rights and enforceability, and instead

claim, intuitively have two features: (1) one agent treats another in a way that the first owes it to the second to treat them, and (2) the second agent owes the first gratitude in response. By virtue of having these two features, these cases represent counterexamples to the Orthodox Thesis.

I then argue that these cases have a further feature in common: part of what the duties in question require of an agent, in context, is to act in such a way that they display a kind of good will to a second, and specifically to act in such a way that they treat the second as an end in themselves, taking the ends of the second as ends of their own. And it is this feature—that the agent acts on a duty that requires them to display good will to another agent—that explains why the second agent owes the first gratitude in response: the first displays good will of the kind that triggers a duty of gratitude. Some moral duties—including certain duties of beneficence, gratitude, and apology—require us to act in ways that display precisely this kind of good will to others. While the Orthodox Thesis may be true when restricted to other duties—in particular, when limited to what some have called "juridical" duties—it is false when asserted in full generality, due to the existence of duties that require us to express good will to one another.

I conclude by addressing an objection to my argument. It appeals to the central premise in an argument commonly given in favor of the Orthodox Thesis, which claims that feeling gratitude involves representing what one is grateful for as something to which one was not normatively entitled. If this premise were true, then the purported counterexamples to the Orthodox Thesis would involve morality requiring agents to represent the moral landscape incorrectly, or requiring agents to ignore the fact that in these cases, they are treated in ways that they are owed. But I argue that we can explain both the intuitive appeal of the claim that feeling gratitude involves representing what one is grateful for as something to which one was not normatively entitled, as well as why this claim is false. My account does not imply that agents are required to represent the moral landscape incorrectly in feeling grateful.

## 1. FOUR COUNTEREXAMPLES TO THE ORTHODOX THESIS

In this section and the next, I will provide a series of cases that, I argue, are counterexamples to the Orthodox Thesis. Each case has two features: (1) one agent treats another in a way that the first owes it to the second to treat them, and (2) the second agent owes the first gratitude in response. In particular, the

use 'claim' as the theory-neutral correlate to directed obligation. For a different argument against the Orthodox Thesis, according to which we owe someone gratitude when they respect our rights in a way that is "notable" or makes them a "moral standout," see McConnell, "Gratitude, Rights, and Moral Standouts"; and Helm, "Gratitude and Norms."

first agent provides a benefit to the second in a way that expresses good will, thereby triggering a duty of gratitude despite the first agent merely doing what is required of them. Afterwards, I will look more closely at what unifies these cases. But first I will provide the series of cases, arguing that each has features 1 and 2.

> *Supermarket*: *Y* is in line for the cashier at the supermarket, and while walking up to the cashier, *Y* trips and drops the cans that they were carrying. *X* is standing behind *Y* in line and notices that *Y* will have a difficult time trying to pick up the cans themselves. Holding only one item themself, *X* picks up the cans and helpfully places them next to *Y*.
>
> *Beach Rescue*: *Y* is swimming in the ocean, gets caught in a riptide, and begins to struggle to stay afloat after fighting against the current. *X* is nearby on a small boat and is trained in water rescue. While rescuing *Y* would no doubt be difficult, *X* is a sufficiently strong swimmer that *X* does not face any significant risk of drowning or serious injury. *X* notices *Y*'s peril and jumps into the water. *X* reaches *Y* before they drown and successfully hauls *Y* back to the boat, saving *Y*'s life.
>
> *Business Competition*: Years ago, *Y* heroically saved *X*'s life, and the two have not encountered one another since. *X* now owns a business and is trying to expand into new markets. *X* is choosing between two areas in which to open a new store, and while they predict the first area to yield marginally higher profits, they also recognize that opening the store there will drive a small store out of business. But while *X* is considering opening the new store, *Y* comes to *X* and informs *X* that *Y* is the owner of the small store, and asks *X* not to open their new store in this area. Out of recognition and appreciation for what *Y* did for them years ago, *X* refrains from opening the store in *Y*'s area.[5]
>
> *Hurtful Joke*: *X* and *Y* are at a party, and the attendees are enjoying each other's company by laughing and telling jokes. Some of these jokes involve making good-natured fun of one another. *X* makes one such joke at *Y*'s expense, but the joke hits a sore spot for *Y*, who becomes quiet and soon leaves the party. While *X* didn't know that *Y* had this particular sore spot, *X* was in a position to know that jokes of this kind can be hurtful and that even when friends make jokes at one another's expense, this type of joke is considered over the line. The next day, after another

5 This case is from Manela, "Obligations of Gratitude and Correlative Rights," who uses it to argue that there are genuine obligations of gratitude.

attendee informs *X* that their joke was hurtful to *Y*, *X* reaches out to *Y* and apologizes. *X* acknowledges that they were inadequately sensitive to the hurt that their joke was liable to cause, sincerely expresses that they value their friendship, and promises to be more sensitive to *Y*'s feelings in the future.

Each case has a few important features. First, in each, *X* provides a kind of help or benefit to *Y*, and does so in a way that expresses good will to *Y*. Importantly, the provision of a benefit from good will is what triggers a duty of gratitude.[6] And this seems to match our intuitions about the cases: *X* provides the kind of help or benefit that calls for gratitude in response. However, contrary to the Orthodox Thesis, *X*'s conduct also seems required: *X* owes it to *Y* to treat *Y* as they do. A defender of the Orthodox Thesis must, then, do one of two things: either claim that *Y* does not actually owe *X* gratitude, or else claim that *X* treats *Y* in a supererogatory rather than required way. In order to forestall both types of response, I will argue in some detail that both feature 1 and feature 2 are present in each case.

I will begin in this section by arguing that feature 1 holds in each case—that is, that in each case, *X* owes it to *Y* to treat *Y* in the way that *X* does. And in order to establish that feature 1 holds in each case, I will first argue that in each case, *X* is required to act as *X* does and will then argue that *X* owes it to *Y* to do so.

In these four cases, we are presented with four different moral duties: in Supermarket, *X* has a duty of (minor) aid or beneficence; in Beach Rescue, *X* has a duty of rescue; in Business Competition, *X* has a duty of gratitude; and in Hurtful Joke, *X* has a duty of apology.[7] Let us take each in turn.

In Supermarket, if *X* fails to help *Y* by picking up the cans, *X* would express a kind of indifference to *Y* that would warrant blame. Especially when it is so easy to help someone who is clearly in need, this kind of indifference involves failing to take account of someone's interests. Of course, if it would be relatively onerous for *X* to provide aid, then failing to pick up the cans would not express this indifference and would similarly fail to warrant blame. But given that it is easy for *X* to help, failing to do so would be *prima facie* blameworthy, indicating

6 For an important early paper that identifies the grounds of gratitude as the provision of a benefit from good will (or "benevolence"), see Berger, "Gratitude." Note that while it is controversial whether a duty of gratitude requires an *actual* or merely an *attempted* benefit, and it is controversial what precise motives are sufficient to trigger a duty of gratitude, it is uncontroversial that duties of gratitude are triggered by the provision of a benefit from good will *in some sense*. See the helpful discussion of these points in Manela, "Gratitude."

7 Depending on how you count, however, there may be three rather than four types of duties in these cases, since the duty of rescue involved in Beach Rescue may be thought to be a special case of the duty of aid or beneficence, which is also involved in Supermarket.

that *X* is required to help.[8] Granted, the stakes in this case are quite low—*Y* will not suffer any great misfortune if *X* does not help by picking up the cans. But this does not show that failing to help would not be wrong; rather, it shows that the wrong would merely be a fairly minor one in the grand scheme of things. Accordingly, *X* is required to help *Y* by picking up the cans.

But not only is *X required* to help *Y* by picking up the cans; further, *X owes* it to *Y* to pick up the cans. That is, *X* would not just act wrongfully by failing to help but, further, would wrong *Y* by doing so. In order to tell whether and to whom some duty is directed, recall the claim about the relation between resentment and the demands of morality described above. This is the claim that *A* is warranted in resenting *B* only if *B* wrongs *A*, i.e., if *B* treats *A* in a way that *B* owes it to *A* not to treat them.[9] Because this claim provides a necessary condition on warranted resentment, it provides us with a test for identifying whether and to whom some duty is owed: if *B* would be warranted in resenting *A* for acting in some way, then *A* owes it to *B* not to act in this way. Accordingly, if *Y* would be warranted in resenting *X* for failing to help by picking up the cans, then *X* owes it to *Y* to pick up the cans. (Call this way of determining whether and to whom some duty is owed the *resentment test*.) And indeed, *Y* would seem to be warranted in resenting *X* for failing to pick up the cans. We wouldn't consider *Y*'s resentment to be misplaced, for in failing to pick up the cans, *X* would show *Y* the type of indifference or disrespect described in the previous paragraph. So not only is *X* required to help *Y* by picking up the cans, but further, *X* owes it to *Y* to help by doing so. Appealing to the resentment test thus confirms that *X* owes it to *Y* to help by picking up the cans.

Before moving to the other cases, I want to preempt two worries about my appeal to whether *Y* would be warranted in resenting *X* for failing to pick up the cans. The first concerns the role and dialectical effectiveness of the resentment test, and the second concerns indifference, ill will, and social expectations.

### *1.1. Resentment and Other Hallmarks of Wronging*

First, in appealing to the resentment test, I infer from the claim that *Y* would be warranted in resenting *X* for failing to pick up the cans (itself justified by appeal

8 *X* is only *prima facie* blameworthy, since *X*'s failure to help could be justified or excused by other factors concerning *X*'s circumstances, knowledge, etc. In what follows, I will simply say that *X* is blameworthy, since we can stipulate that in none of the four cases would *X*'s failure to act be justified or excused by other factors.

9 We can modify this necessary condition on warranted resentment into a necessary and sufficient condition on warranted resentment by adding a clause to this claim: *A* is warranted in resenting *B* only if *B* wrongs *A*, i.e., if *B* treats *A* in a way that *B* owes it to *A* not to treat them, absent excuse or special justification.

to intuition) that *X* owes it to *Y* to pick up the cans. But one may worry about relying on the resentment test in this way, since the connection between resentment and obligation described by the resentment test is itself both substantive and controversial. If defenders of the Orthodox Thesis do not antecedently accept the resentment test, what reason do they have to accept that *X* owes it to *Y* to pick up the cans? Further, this worry takes on added significance in virtue of my argument in the next section that in each of the four cases, *Y* owes *X* gratitude in response. In short, I there use the resentment test to argue that in each of the four cases, *Y* owes *X* gratitude for treating them in a way *Y* is owed, and thus that the Orthodox Thesis is false. But a defender of the Orthodox Thesis may use the same sort of reasoning in the other direction: on the basis of the Orthodox Thesis, they may infer from the fact that *Y* owes *X* gratitude in each case that *X* must not have owed it to *Y* to treat *Y* as they do, and thus that the connection between resentment and obligation described by the resentment test is false. This objection, in sum, suggests that one can reason from the Orthodox Thesis to the falsity of the connection between resentment and obligation described by the resentment test just as easily as one can reason from the resentment test to the falsity of the Orthodox Thesis.[10]

In response, I will briefly note some of the main points in favor of the connection between resentment and obligation described by the resentment test, before describing how my argument can be modified so as not to rely on the resentment test at all. Recall that the resentment test holds that *B* is warranted in resenting *A* only if *A* wrongs *B*, i.e., if *A* treats *B* in a way that *A* owes it to *B* not to treat them.[11] The basic reasoning behind this claim concerns the connections between resentment, ill will, treating someone with proper regard, and wronging. We can provide an argument for the connection between resentment and obligation described by the resentment test as follows:

1. *B* is warranted in resenting *A* only if *A* displays ill will toward *B*.
2. *A* displays ill will toward *B* just in case *A* fails to treat *B* with proper regard.
3. *A* fails to treat *B* with proper regard just in case *A* wrongs *B*.
4. Therefore, *B* is warranted in resenting *A* only if *A* wrongs *B*.

10 Thank you to an anonymous reviewer for raising this objection.

11 Note that the resentment test relies only on a necessary condition for warranted resentment, not a sufficient condition. Just because *A* treats *B* in a way that *A* owes it to *B* not to treat them, *B* would not necessarily be warranted in resenting *A*. *A* could, for instance, have a good excuse for treating *B* in this way, or it could be hypocritical for *B* to resent *A* for treating them in this way.

This argument provides at least *prima facie* support for the resentment test. Its premises are intuitively plausible and entail the conclusion. Indeed, some have argued that its premises express conceptual truths about reactive emotions like resentment and their relation to moral obligations and accordingly that the content of and conditions of justification for resentment cannot be understood independently of the notion of treating others in accordance with the demands of morality.[12] Nevertheless, both the resentment test and this argument in favor of it are controversial, and much more would need to be said to adequately establish the connection between resentment and obligation described by the resentment test. Thankfully, my argument can be modified so as not to rely on the resentment test at all. While the resentment test provides perhaps the most direct method for establishing that in each case, *X* owes it to *Y* to treat *Y* as they do, we can establish this fact in a different way, avoiding reliance on the resentment test.

In particular, for each of the four cases, we can identify other hallmarks or identifiers of directed duties, thus sidestepping issues about the precise relation between resentment and wronging. There are two main alternate identifiers for directed duties that are present in each case. First, in each case, *Y* alone has the standing to remonstrate or complain if *X* does not comply with their duty. And *Y* has the standing to remonstrate against *X*'s noncompliance only if *X*'s duty is directed toward *Y*. Second, in each case, if *X*'s noncompliance triggers duties of apology or repair, these duties would be directed toward *Y*. And *Y* is owed a duty of apology or repair by *X* only if *X* wrongs *Y*.[13] I will first explain why both the standing to remonstrate and being owed apology or repair are tied to being the claimholder of a directed duty and then argue that each is present in the Supermarket case.

Let us first consider the relation between the standing to remonstrate and directed duties. To say that *Y* has the standing to remonstrate with *X* is to say that *Y* has the standing to attempt to influence *X* by citing normative reasons

12 See especially Wallace, *Responsibility and the Moral Sentiments*, ch. 2.

13 Arguably, there is a third alternate identifier we could appeal to: in each case, *Y*'s interests ground *X*'s duty, and according to the interest theory of directed duties, *Y*'s interests ground *X*'s duty just in case *X*'s duty is directed toward *Y*. Although the interest theory delivers the right verdict in each case about whether and to whom *X*'s duties are owed, I will not lean on it as an identifier for directed duties since it is even more contentious than the resentment test. In particular, its main opponent is the will theory, which holds that *X*'s duty is directed toward *Y* just in case whether *X* is obligated is dependent on *Y*'s will—that is, *X*'s duty is directed toward *Y* just in case *Y* has the power to waive *X*'s duty. And the will theory does not return the right verdict on the cases presented here, since duties of gratitude have notably been argued not to provide those to whom they are owed with the power of waiver. See Herman, "Being Helped and Being Grateful."

that *X* already possesses but that may be motivationally silent to them. For example, *Y* might remonstrate by saying such things as "Are you seriously just going to stand there?" or "You know, I could use a little help here." By remonstrating, *Y* would attempt to exert more force on *X* than by merely requesting *X*'s help.[14] Importantly for my purposes, not just anyone has the standing to remonstrate with someone about their noncompliance with some obligation. If I notice that you are not complying with a promise you made to a third party, I might remind you of the promise or describe how the third party might feel when they learn of your noncompliance. But I lack the standing to remonstrate with you about your noncompliance. In particular, only the person to whom your duty is directed has the standing to remonstrate with you about your noncompliance. That is, *Y* has the standing to remonstrate with *X* about whether *X* $\phi$s only if *X* owes it to *Y* that they $\phi$.[15]

Next, let us consider the relation between directed duties and duties of apology and repair. Here the connection is even more straightforward than with the standing to remonstrate. Owing someone an apology or some other form of repair such as compensation is explained by having wronged them or by having violated a duty that was owed to them. When we wrong someone, we can sometimes do harm to third parties. For instance, suppose that I promise to give you some apples, and you lead a third party to believe that you will give them the apples so that they can bake an apple pie. If I break my promise to you, I set back both your interests and the third party's interests. But my subsequent duties of apology and repair pertain only to you, not to the third party. And this is because being owed duties of apology and repair coincides with being the claimholder of a directed duty. More specifically, *A* owes *B* duties of apology and repair only if *A* wrongs *B* or if *A* fails to comply with a directed duty owed to *B*. Accordingly, the standing to remonstrate and duties of apology and repair stand as apt alternative identifiers for being the claimholder of a directed duty,

14 For more on the standing to remonstrate, as well as its connection to "imperfect rights," see Manela, "Obligations of Gratitude and Correlative Rights."

15 Often, the individual to whom some duty is directed has not only the standing to remonstrate but further the standing to *demand*. Like remonstrating, demanding involves an attempt to bring about someone's compliance with a duty, but demanding is more forceful than remonstrating and constitutes an attempt to *enforce* one's claim. But we cannot appeal to the standing to demand as an identifier of directed duties in the present context, since duties like gratitude and apology notoriously do not provide their claimholders with the standing to demand. On the relation between the standing to remonstrate and the standing to demand, see again Manela, "Obligations of Gratitude and Correlative Rights"; and for an account of why duties of gratitude do not provide their claimholders with the standing to demand, see Segal, "Gratitude and Demand."

independently of any claims about the connection between directed duties and resentment.

Finally, in Supermarket, we can confirm that *X* owes it to *Y* to pick up the cans by pointing to the standing to remonstrate and duties of apology and repair. As noted above, it seems that *Y* has the standing to remonstrate with *X* about *X*'s picking up the cans. Although it would seemingly be inappropriate for *Y* to launch into a full tirade in order to pressure *X* into picking up the cans, it would be appropriate for *Y* to cite the reasons why *X* ought to help by picking up the cans, in an attempt to get *X* to pick up the cans—for instance, by citing the fact that they could use a bit of help or the uncaringness of simply standing by while *Y* struggles to pick up the cans. And further, if *X* does stand by without helping, it seems that *X* would owe *Y* an apology. Given that *X* and *Y* have no personal relationship and that the stakes of the aid are quite low, *X* need not do much more than a simple verbal apology—something along the lines of "I'm sorry I didn't help you just then; I was wrapped up with going about my own day, but I shouldn't have ignored your situation." Given the low stakes of the case, it would be inappropriate for *Y* to remonstrate at great length or with serious anger, and if *X* does not help, *X* would not owe *Y* a very extensive apology or other form of repair. Nevertheless, *Y* does have the standing to remonstrate, and *X* would owe *Y* an apology if *X* fails to help. Since the standing to remonstrate and being owed duties of apology or repair serve as alternate identifiers of directed duties, we can thus establish that *X* would wrong *Y* by failing to help—without reliance on the resentment test.

The final point worth mentioning regarding this way of modifying my argument so as to avoid relying on the resentment test is that just as defenders of the Orthodox Thesis might deny the connection between resentment and wronging expressed by the resentment test, they might also deny the connections between the standing to remonstrate, duties of apology and repair, and directed duties that I have just argued for. Each of these connections represents a substantive claim about the nature of directed duties, and it is theoretically open to defenders of the Orthodox Thesis to take issue with any of them. But in order to deny my claim that *X* owes it to *Y* to pick up the cans (as well as my parallel claims for the other three cases), they would have to reject nearly all of the apparent identifiers of directed duties and would be left with a deeply controversial view of how to identify whether and to whom a duty is owed. So while it is open to defenders of the Orthodox Thesis to reject not only the resentment test but also the alternate identifiers of the standing to remonstrate and being owed duties of apology and repair, doing so represents biting a sufficiently large bullet that I take myself to have put significant pressure on defenders of the Orthodox Thesis who wish to deny that in each or all of the four cases, *X* owes it to *Y* to treat them as they do.

*1.2. Indifference, Ill Will, and Social Expectations*

The second worry worth discussing before proceeding to the other three cases concerns the kind of indifference that *X* would display to *Y* if *X* failed to help by picking up the cans. I claimed above that if *X* failed to pick up the cans, *X* would display a type of indifference to *Y* that, in the context of Supermarket, constitutes a display of ill will to *Y*. And because resentment is an appropriate response to ill will, *Y* would be warranted in resenting *X* for failing to help. Finally, because *Y* would be warranted in resenting *X*, I concluded that *X* owes it to *Y* to help by picking up the cans. However, one might wonder why, exactly, *Y* would be warranted in resenting *X*'s indifference, and correspondingly, why *X* owes it to *Y* to help. Importantly, we are not subject to a blanket moral prohibition on being indifferent to others. We are not morally required to spring into action whenever we see someone who we can help to complete a minor task. Suppose, for instance, that from across the street, I notice you struggling to open a bottle of water. Not only am I not required to cross the street to help you open it; you might reasonably find it strange or uncomfortable for me to approach you out of the blue to help. Refraining from helping you to open the water bottle involves a type of indifference—but a perfectly innocent type. Why should we think that helping in Supermarket is different from helping you open the bottle of water? That is, why should we think that indifference to a stranger is permissible in one case but impermissible in another?[16]

The answer lies in the presence of social expectations of a particular type. When *X* and *Y* share the right kind of expectations about when and how individuals should help one another, and *X*'s refraining from helping *Y* would violate these expectations, then the indifference expressed by refraining from helping would constitute ill will rather than merely "innocent" indifference. I will first describe the relevant type of social expectations in more detail, then explain how they derive from associated conventional norms, and finally outline the considerations that give these conventional norms moral force.

The social expectations relevant to the question of when indifference rises to the level of ill will are expectations concerning: (1) when and from whom one will receive help; and (2) the ways in which one will be held accountable for helping or failing to help others. We carry these expectations in the background of many or most of the social interactions we have: for instance, we expect (if only implicitly) that if someone sees something fall out of our pocket, they will let us know or pick it up and hand it to us. And we expect that if we violate others' expectations, we may be held accountable through reactive emotions like blame or resentment. In the absence of any such expectations to help, indifference to

16 Thank you to an anonymous reviewer for pressing me on this question.

others typically does not constitute ill will. My indifference to your difficulty in opening the bottle of water, for example, seems not to constitute ill will in part because you have no expectation that someone in my position would come to your aid. My indifference to someone rises to the level of ill will when there is an expectation that, in the circumstances, I will help them. When such expectations are present, refraining from helping is not mere indifference but a knowing violation of another person's expectation that I help them.

However, not just any such expectations seem capable of making indifference constitute ill will. Suppose that you expect others to hold doors open for you if you are within one hundred feet of a door (and also expect to be held accountable yourself for not doing so for others). Other people would presumably violate your expectation on a routine basis, but they would not thereby express ill will toward you—even if you might feel as though they do. This is because as a society, we have settled on a conventional norm of holding doors open for others only when they are (roughly) immediately behind us. The fact that people in general have very different expectations from you about whether individuals will or ought to hold doors open for others who are relatively far away means that violating your idiosyncratic expectations does not constitute ill will. Whether one individual's indifference toward another rises to the level of ill will is partly a function of whether the indifference violates the other's expectations about how they will be helped, but not just any expectations will do. Indifference toward someone constitutes ill will when it violates their expectations about how they will be helped, where these expectations are derived from generally accepted conventional norms about when and how individuals should help one another.

The ability of these conventional norms to determine when indifference is innocent and when it constitutes ill will depends on these norms having some degree of moral force. If they were strictly nonmoral norms, they could give rise to expectations that could variably be satisfied or violated by others' conduct, but violations of them would not constitute ill will. That is, if they were strictly nonmoral norms, then violations of the expectations they give rise to would not be morally blameworthy, would not justify resentment, and would not ground obligations to help one another, as in Supermarket. Accordingly, in order to explain how conventional norms can determine whether indifference rises to the level of ill will, we need to explain how these norms can take on moral force. When and why do the conventional norms that give rise to expectations concerning when and how to help acquire moral force?

Although we could explain this moral force in a number of ways, one promising route holds that conventional norms concerning when and how to help acquire moral force when and because their general acceptance solves a

certain type of coordination problem. In the absence of the general acceptance of conventional norms concerning when and how to help, individuals would be unable to rely on others helping them in any specific way or in any specific circumstances. This is because different individuals have widely divergent preferences concerning the ways in which they would like to be helped, the ways they are inclined to help others, and the amounts of effort they believe that individuals should exert to help one another. But at the same time, because none of us can avoid needing help from others in order to achieve our ends (at least from time to time), it is better from the standpoint of each individual to live in a community that has adopted *some* set of conventional norms rather than none, even if the conventional norms accepted by the community do not precisely match their own conception of how individuals should help one another. Because individuals therefore benefit from living in a community that generally accepts conventional norms concerning how to help one another, they can be justifiably held accountable in terms of these norms with respect to whether or not they help in particular circumstances.[17]

To summarize: indifference to someone rises to the level of ill will when and because it violates a social expectation derived from a conventional norm concerning when and how to help one another. These conventional norms have moral force when and because their general acceptance provides a solution to a coordination problem that would otherwise occur. In the example of you struggling to open a water bottle, there is no generally accepted conventional norm requiring individuals to cross the street to help. But in Supermarket, there is a conventional norm that requires individuals to help when they are in the immediate vicinity of someone who drops some items and needs some help (at least when it is relatively easy to do so).[18] Insofar as *X* and *Y* are both members of the social practice that generally accepts this norm, they share expectations about the circumstances in which individuals should help one another. If *X* refrains from helping, then *X* violates *Y*'s expectation that *X* helps, thus expressing not only indifference to *Y* but ill will to *Y* as well.

17 This account of the source of the moral force of conventional norms concerning how to help one another is here presented only in schematic form. For an argument that appeals to conventional norms' ability to solve this type of coordination problem to justify holding one another accountable to moral norms more broadly, see Gaus, "The Demands of Impartiality and the Evolution of Morality." And for a related but distinct explanation of the moral force of these conventional norms in terms of respect rather than coordination, see Stohr, *On Manners*.

18 If you have doubts about how widely accepted this norm is, suppose that Supermarket takes place in an area where politeness, friendliness, and courtesy are strongly held social values—like many small towns in the American Midwest.

*1.3. Returning to the Cases*

Let us turn back to the other cases beside Supermarket. In Beach Rescue, if *X* fails to help *Y* by jumping into the water and attempting to save *Y*, *X* would express a similar kind of indifference to *Y* as in Supermarket—but with much higher stakes. Although the kind of help that *Y* needs in Beach Rescue is much more onerous than the help involved in Supermarket, this would provide no justification for failing to help, since *Y*'s life is at stake. The fact that *Y*'s life is at stake shows that failing to help would be wrong, at least so long as *X* would not be risking their own life in the process.[19] So *X* is required to help *Y* by jumping into the water and attempting to save them. And further, the same test that we used in Supermarket indicates that *X* owes it to *Y* to try to save their life: if *X* were to stand idly by, then *Y* would be warranted in resenting *X*. Of course, if *X* were to stand idly by, then *Y* would likely perish. But the relevant question is not whether *Y* would have the chance to resent *X* but whether such resentment would be warranted. And in Beach Rescue, *X* failing to help would express a more extreme form of the kind of indifference involved in the failure to help in Supermarket. So *X* owes it to *Y* to jump into the water and attempt to save them.

In Business Competition, if *X* does not accede to *Y*'s request and opens the new store in *Y*'s area anyway, then *X* would express a lack of recognition and appreciation for *Y*'s lifesaving aid. Here and now, *X* has the opportunity to express their recognition and appreciation—in short, their gratitude—for this aid, and failing to do so would express ingratitude. Saving someone's life is such a significant benefit that, at least typically, it triggers a duty of gratitude for the person saved. And while we often have considerable latitude in determining just how to express gratitude to those who benefit us, in Business Competition,

19 One might wonder how much risk one is required to incur in order to save someone's life: surely, saving someone's life is required when doing so would take only minimal effort, and, on the other hand, we seem not to be required to sacrifice our own lives in order to save someone else. This is a difficult question even when all else is equal, and it is made more complex still when we consider other complicating factors that may matter, such as how someone came to need rescue, whether the potential rescuer has led others to rely on their willingness or ability to rescue, and the fairness of requiring individuals in the potential rescuer's position to incur the relevant risks. As a rough guideline, it seems that an individual is required to incur risks in order to save someone's life when: (1) the probability of serious harms (e.g., contracting a monthslong illness) is quite low; and (2) any harms with a significant chance of occurring (e.g., a greater than 10 percent chance) are relatively minor. Of course, even this rough guideline is not on its own enough to settle difficult borderline cases. But in Beach Rescue, because *X* is a sufficiently strong swimmer and is trained in water rescue, the risks are low enough to conclude that *X* is required to (attempt to) save *Y*.

this ordinary latitude is absent.[20] *Y* saved *X*'s life, and *X* now has the opportunity to save *Y*'s livelihood—or else to eliminate it. Insofar as opening the new store despite *Y*'s request would express ingratitude, and *Y*'s earlier lifesaving aid triggers a duty of gratitude for *X*, *X* is required to refrain from opening the new store in *Y*'s area. Further, the resentment test for determining whether and to whom a duty is directed has the result that *X* owes it to *Y* to refrain from opening the new store and would wrong *Y* by failing to do so. If *X* were to open the new store and drive *Y* out of business, *Y* might reasonably resent *X*, thinking something along the lines of "After all I did for *X*, this is the thanks I get?" Accordingly, *X* not only is required to refrain from opening the new store in *Y*'s area but in fact owes it to *Y* to do so.

One might wonder, however, whether the latitude that duties of gratitude typically provide is really absent in this case. Ordinarily, duties of gratitude allow agents to express gratitude in a variety of ways. Suppose that my car breaks down, stranding me on the side of the road, and you come to pick me up in the middle of the night. All else equal, your assistance is sufficient to trigger a duty of gratitude on my part. But this duty does not require me to express my gratitude in any particular way. Surely a verbal expression of appreciation is a good start, but beyond that, I might buy you dinner or offer to help you with a home renovation project or something else. Part of what makes an action able to express our sincere gratitude rather than our mere willingness to repay a transactional debt is the fact that we perform it freely or of our own accord. And to the extent that an action's being free in this sense is at odds with rigoristic rules about precisely how to express gratitude, we can see why duties of gratitude provide latitude in a way that many other duties do not. Why, then, should we think that this typical latitude is absent in Business Competition? That is, why not think that *X* could express their gratitude to *Y* in some way other than refraining from opening the new store in *Y*'s area?

Without defending a full account of the latitude involved in duties of gratitude (or in "imperfect duties" more generally), there are a few important features of Business Competition that make it different from other cases featuring duties of gratitude. First, the original benefit that *Y* provided to *X*—saving *X*'s life—is significantly larger than most benefits. While the magnitude of the benefit seemingly cannot on its own eliminate the latitude provided by a beneficiary's duty of gratitude, it does mean that the beneficiary's expression of gratitude must also be significant. (A casual "thank you" suffices to express gratitude when someone holds a door open for us, but not when someone saves our life.)

20 I stay neutral here on what feature of Business Competition—or of duties of gratitude more generally—explains the fact that the typical latitude involved in duties of gratitude is absent here.

Second, *X* and *Y* do not have an ongoing relationship, and *X* did not have the opportunity to express their gratitude to *Y* at any earlier point—while they may have thanked *Y* for saving them at the time, refraining from opening the new store in *Y*'s area may well be their only chance to reciprocate *Y*'s earlier benefit. Third, the cost to *X* of refraining from opening the new store in *Y*'s area pales in comparison with the benefit to *Y* of doing so. *X* only anticipates marginally better profits from opening the new store in *Y*'s area rather than another area, but *Y* would lose their business and livelihood unless *X* refrains from doing so. Fourth, *Y* requests that *X* refrain from opening the new store in *Y*'s area. While requests concerning how a beneficiary expresses their gratitude do not (at least ordinarily) make it obligatory for a benefactor to express gratitude in the specific way requested, they do provide additional reason in favor of expressing gratitude in that way rather than others—at least so long as the request is made in good faith and without making the tenor of the interaction transactional.

These four factors—the magnitude of the benefit, limitations on the beneficiary's opportunities to express gratitude, the ratio of costs to benefits, and the benefactor's request—each constrains the degree of latitude that a beneficiary has with respect to how to express their gratitude. And when each is present, as in Business Competition, they can constrain the latitude typically provided by duties of gratitude to the point of eliminating it altogether. Ordinarily, duties of gratitude allow agents to determine for themselves which specific ways to express gratitude. But when these constraining factors are present, there can be fewer actions that can express sincere and appropriate gratitude, to the point that sometimes there is only one such action. In cases like Business Competition, *X* cannot choose alternative means of expressing their gratitude—sending *Y* flowers, or even writing *Y* a check, would not demonstrate that *X* genuinely appreciates *Y*'s original rescue and wants to reciprocate it. Insofar as duties of gratitude require us to express our appreciation and (when possible) reciprocate benefits provided to us, the constraining factors can limit the extent of our latitude in doing so.[21] Accordingly, *X* owes it to *Y* to refrain from opening the new store in *Y*'s area.

21 It is worth noting that even cases in which a duty of gratitude *does* provide latitude can arguably play the same role in my argument that I claim Business Competition does. For even in such cases, a beneficiary can act in a morally required way—namely, expressing gratitude—and a benefactor can owe the beneficiary gratitude in response. And because in such cases, one agent treats another in a way that they owe them to, and the second owes the first gratitude in response, such cases would still represent counterexamples to the Orthodox Thesis. The primary difference for the purposes of my argument between such cases and those that, like Business Competition, lack the latitude typically provided by duties of gratitude concerns the level of description under which a beneficiary's action is morally required. In cases without latitude, the beneficiary's (that is, *X*'s) action is required

Finally, in Hurtful Joke, *X* inadvertently hurts *Y*'s feelings by making a joke that hits a sore spot for *Y*. To be fair, inadvertently hurting someone's feelings is no grave moral sin—it is closer to a casualty of living in a community of people who each have distinct sensibilities and vulnerabilities, making it close to inevitable that we step on one another's toes from time to time. Nonetheless, if *X* were to refuse to apologize to *Y*, then *X* would seem to express disrespect to *Y*: refusing to apologize would demonstrate that *X* does not consider *Y*'s interest in emotional well-being and feeling secure in their group of friends to be weighty enough to warrant apologizing. Further, in refusing to apologize, *X* would signal that they will not take steps to avoid hurting *Y*'s feelings again in the future. So even though we may not be inclined to blame *X* for inadvertently hurting *Y*'s feelings in the first place (or at least we may not be inclined to blame *X* very much), it does seem that *X* is required to apologize for doing so. Additionally, the resentment test yields the same result as in the previous three cases: if *X* refuses to apologize, it seems that *Y* would be warranted in resenting them. *Y* might reasonably think to themselves, "I'm sure that *X* didn't mean it, but still—doesn't it matter to them that the joke was hurtful?" As with the first three cases, then, not only is *X* required to treat *Y* in the way that *X* does; further, *X* owes it to *Y* treat them in this way.

All four cases thus have feature 1: one agent treats another in a way that the first owes it to the second to treat them. I have gone into considerable detail in arguing that for each case, *X* owes it to *Y* to treat *Y* as *X* does, in order to prevent a defender of the Orthodox Thesis from objecting to my argument on the grounds that these are cases of mere supererogation and so are consistent with their view. But in order to serve as counterexamples to the Orthodox Thesis, these cases must also have feature 2: the second agent owes the first gratitude in response.

## 2. GRATITUDE AND OBLIGATION

I will now argue that each case also has feature 2: *Y* owes *X* gratitude for acting as *X* does. In each case, *X* provides a benefit to *Y*, and does so in a way that expresses good will to *Y*.[22] And since the provision of benefits from good will

under the description of the specific action performed—in this case, "refraining from opening the store in *Y*'s area." By contrast, in cases with latitude, the beneficiary's action is required under the more general description of "expressing gratitude." But insofar as in both types of cases, one person's morally required expression of gratitude triggers a duty of gratitude on the other person's part, both types of cases provide counterexamples to the Orthodox Thesis.

22 The fact that *X* provides a benefit to *Y* is perhaps least straightforward in Hurtful Joke. But I take it that an apology can constitute a benefit at least when it helps to mend a damaged relationship, insofar as the relationship is valuable to each person.

triggers a duty of gratitude, it follows that *Y* owes *X* gratitude in response—despite the fact that *X* owes it to *Y* to treat *Y* as they do. But because this fact makes the four cases counterexamples to the Orthodox Thesis, it is worth finding extra confirmation of the fact that *Y* owes *X* gratitude in response. In particular, defenders of the Orthodox Thesis might try to save their view by arguing that in each case, it would be praiseworthy but supererogatory for *Y* to express gratitude. By contrast, I am claiming that *Y* owes it to *X* to express gratitude, and so *Y*'s gratitude is required, not supererogatory.

In saying that *Y* owes *X* gratitude in response, I mean that *Y* owes it to *X* to *express* gratitude, not just to feel gratitude. Expressions of gratitude, at least in the sense I mean, are primarily actions that someone performs out of recognition and appreciation of what they are grateful for, rather than verbal expressions that inform someone that they feel grateful.[23] In order to defend against the worry that gratitude in these cases would be praiseworthy but more generous than morality requires, for each case, I will argue first that gratitude is an *appropriate* way for *Y* to respond to *X*'s conduct and then that *Y*'s gratitude is not merely appropriate but in fact *owed* to *X*. Let us turn back to the four cases.

In Supermarket and Beach Rescue, *X* provides two types of help or aid to *Y*—in the former, the aid is quite minor, while in the latter, the aid is vital. And *X* helps *Y* without being externally forced or coerced to do so. Not only does *X* help *Y* in both cases; *X* does so of their own accord. And in doing so, *X* displays to *Y* a kind of good will: *X* wants to help *Y*, and (let us say) not simply in order to get something from *Y* in return. Further, suppose that in each case, following *X*'s help, *Y* both feels and expresses gratitude to *X* for the help.[24] Would such gratitude strike us as inappropriate or unfitting? I do not think so—I do not think that many people would, in *X*'s position, find *Y*'s gratitude odd or inappropriate. *X* helps *Y* and exhibits a kind of good will in doing so. In such circumstances, gratitude is a natural response.

23 Sometimes a verbal expression is sufficient to fulfill a duty of gratitude, but I am primarily interested in the sense in which we can owe others gratitude in the form of actions that reciprocate what one is grateful for. Further, there is plausibly a sincerity condition on expressions of gratitude: an action is prevented from expressing gratitude if the agent actually feels ungrateful. Still, the locus of "expressions of gratitude" as I use the phrase is action, not speech or feeling.

24 I will talk of both feeling and expressing gratitude in order to avoid the question of what exactly duties of gratitude require of us. I elsewhere argue that they should be understood as duties to act in ways that express gratitude, where "expressing gratitude" is both determined by conventional understandings of what types of behavior count in context as expressing gratitude, and subject to a sincerity condition that rules out the possibility of expressing gratitude while feeling ungrateful.

But further, *Y*'s gratitude is not merely *appropriate*; it is something that *Y* *owes* to *X*, in the sense that *Y* would wrong *X* by failing to feel or express gratitude in response. *X*'s help triggers a duty of gratitude for *Y*. Why think that this is so? To answer this question, we can turn again to the test concerning resentment: if *A* would be warranted in resenting *B* for failing to feel or express gratitude for *A*'s help, then *B* owes *A* gratitude for *A*'s help. And it does seem that were *Y*'s gratitude not forthcoming, *X* would be warranted in resenting *Y*.[25] Take Supermarket first: if *Y* does not even acknowledge *X*'s help, then it would seem warranted for *X* to resent *Y*. Admittedly, *Y*'s ingratitude in this case certainly would not warrant anything like a longstanding grudge—after all, the help only involves picking up a few cans. But some degree of resentment, perhaps proportional to the relatively minor significance of the interaction, does seem warranted. Next, take Beach Rescue: if *Y* does not thank *X* right after being saved, this seems reasonable, since *Y* would presumably be in a state of shock. But if *Y* has the opportunity to express gratitude after the shock has subsided, *X* might reasonably feel resentful of *Y*'s ingratitude. (After all, they saved *Y*'s life!) Accordingly, in Supermarket and Beach Rescue, *Y* owes *X* gratitude for *X*'s treatment of them.

Next, consider Business Competition. Here again it seems appropriate for *Y* to feel and express gratitude to *X* for refraining from opening the new store in *Y*'s area. In a sense, *X* does not have to accede to *Y*'s request: expanding the reach of one's business is fair game, so far as the competitive market is concerned. And *X* refrains from opening the new store of their own accord, rather than in response to *Y* making a demand that *X* do so, for instance, or because of coercion from some regulatory institution. *X* refrains from opening the new store in order to reciprocate *Y*'s aid years before and to express their appreciation for that aid, thereby expressing good will toward *Y*. In response, then, it is perfectly appropriate for *Y* to feel and express gratitude for *X*'s refraining from opening the new store. Further, *Y* owes such gratitude to *X*: if, once *X* decides to refrain from opening the new store and informs *Y* of this fact, *Y* neither feels nor expresses gratitude, then *X* would be warranted in resenting *Y*. *X* went out of their way to refrain from engaging in an ordinary and profitable business activity, and did so for *Y*'s sake and at *Y*'s request. If *Y* neither feels nor expresses gratitude in return, *X* might reasonably feel taken advantage of. And because *X* would be warranted in resenting *Y* for their ingratitude, we can infer that *Y* owes *X* gratitude for refraining from opening the new store.

25 Note that in all four cases, there seems to be a shared set of social expectations concerning the ways in which individuals are supposed to help one another. It is partially in virtue of both *X* and *Y* sharing these expectations that it seems warranted for *Y* to resent *X* if *X* does not help, as well as for *X* to resent *Y* if *Y* is subsequently ungrateful.

Finally, consider Hurtful Joke. Once again, it seems appropriate for *Y* to feel and express gratitude to *X* for apologizing. *X*'s joke, although hurtful, was not motivated by malicious intent, and *Y* might reasonably think that had *X* known that the joke would hit on a sore spot for *Y*, *X* would not have made the joke. Further, we can suppose that *X*'s apology did not stem from pressure from others to apologize, nor from *Y* demanding that *X* apologize—it was something that *X* decided to do of their own accord, from feeling guilty or otherwise negatively about hurting *Y*'s feelings. *X*'s apology serves to signal that *X* cares about their relationship with *Y* and takes considerations concerning *Y*'s happiness to constrain *X*'s own behavior. In apologizing, then, *X* displays good will to *Y*, making it appropriate for *Y* to feel and express gratitude for the apology in return. Indeed, *Y* might reasonably express this gratitude by forgiving *X* or even by insisting that there is nothing to forgive *X* for. And once more using the resentment test, we can see that *Y*'s gratitude is not only appropriate but genuinely owed to *X*. Following *X*'s apology, if *Y* does not feel or express gratitude, *X* would be warranted in resenting *Y*, at least to some degree. *X* might reasonably feel as though their attempt to repair the relationship and express good will had fallen on deaf ears. "I told *Y* that I wouldn't have made the joke if I had known that it would be hurtful—shouldn't that matter to them?" Of course, resentment may be out of place if *Y*'s lack of gratitude stems from the fact that their feelings are still hurt or from the fact that *Y* feels that *X* should have known better. But supposing that *Y*'s feelings are no longer hurt and that *Y* understands that *X* didn't mean to hurt *Y*'s feelings, if *Y* were not to feel or express gratitude in response to *X*'s apology, then it would be reasonable for *X* to resent *Y*. Accordingly, *Y* owes *X* gratitude for *X*'s apology.

One might worry, however, that the plausibility of the claim that *Y* owes *X* gratitude for *X*'s apology rests on the implicit assumption that forgiving someone is a way of expressing gratitude to them. According to this thought, what *Y* first and foremost owes *X* is forgiveness, not gratitude, and it is plausible to claim that *Y* owes *X* gratitude only insofar as forgiving *X* is a way of expressing gratitude to *X*. This would be a serious difficulty for my analysis of Hurtful Joke, since this gratitude-centric view of forgiveness is at best quite controversial and at worst a straightforwardly false account of forgiveness. However, we need not accept any such analysis of forgiveness in order to accept the claim that *Y* owes *X* gratitude for *X*'s apology. This is because though forgiveness in light of an apology and gratitude for the apology itself often go hand in hand, gratitude is neither necessary nor sufficient for forgiveness.

First, note that we can forgive someone without them first apologizing. Our ability to do so shows that gratitude is not necessary for forgiveness. And even when we forgive someone who has apologized, our forgiveness need not

involve gratitude for their apology—consider an apology that we suspect is not genuine, but we forgive the person nonetheless. Second, we can be grateful for an apology without thereby forgiving. If someone has deeply hurt my feelings, I might be grateful for an apology without thereby feeling ready to forgive them. (We can, then, continue to resent someone for wronging us in the first place, even while we are grateful for their efforts to make up for their wrongdoing.) Gratitude is thus neither necessary nor sufficient for forgiveness.

Accordingly, when I claim that in Hurtful Joke, *Y* owes *X* gratitude for *X*'s apology, I mean to be ambivalent about whether *Y* owes *X* forgiveness. Whether we are obligated to forgive—and indeed whether we are able to forgive—is sensitive to different emotions from whether we are obligated to express gratitude. For while we can be grateful for an apology while continuing to be hurt by or angry about the wrong done to us, it is much more difficult (if not impossible) to forgive while retaining our hurt or anger. Given that *Y*'s feelings were quite hurt, *X*'s apology might not be enough to make it obligatory for *Y* to forgive *X*. But given that *X*'s apology was sincere and that the emotional pain that *X* caused *Y* was inadvertent, *Y* does at least owe *X* gratitude for apologizing. My claim that *Y* owes *X* gratitude for *X*'s apology thus does not rest on the controversial, if not wholly implausible, view that forgiveness is a form of gratitude.

All four cases thus have both features 1 and 2: one agent treats another in a way that the first owes it to the second to treat them, and the second owes the first gratitude in response. All four cases are thus counterexamples to the Orthodox Thesis, which holds that *A* never owes *B* gratitude for *B*'s treating *A* in a way that *B* owes it to *A* to treat them. The four cases feature a number of different moral duties, with the aim of putting to the side concerns that might arise about specific cases—e.g., about whether *Y* really owes *X* gratitude for *X*'s apology in Hurtful Joke or about whether *X* really owes it to *Y* to pick up the cans in Supermarket. So long as we find at least one case that has both features 1 and 2, the Orthodox Thesis is false. Nonetheless, I think that all four cases are counterexamples and that there is a common feature that explains why cases of this kind are apt to function as counterexamples to the Orthodox Thesis. In particular, I think that the duties involved in these cases are unlike many other moral duties and have a special connection to the quality of will expressed in fulfilling them. I turn now to this further feature at issue in the four cases.

## 3. DUTIES OF GOOD WILL

What explains why *X*'s duty-fulfilling actions in the four cases presented above trigger duties of gratitude for *Y*? Duty-fulfilling actions do not in general have this property: I do not owe you gratitude for respecting my right to bodily

autonomy, for refraining from deceiving me, or for treating me in countless other ways that you owe me.[26] Why are the cases above different? The answer that I will argue for in this section is that part of what the duties involved in these cases require of *X* is that *X* acts in a way that expresses good will to *Y*. The duties at issue are what we can call *duties of good will*. In treating *Y* in the way that *Y* is owed, then, *X* expresses good will to *Y*. And it is this fact—that *X* expresses good will to *Y* in treating *Y* in the way that *Y* is owed—that explains why *Y* owes *X* gratitude in response. I will first go into more detail concerning what it takes to express good will in the relevant sense and argue that the duties involved in the four cases are duties of good will. I will then argue that this feature of the four cases is what explains why *Y* owes *X* gratitude in each, despite the fact that *X* treats *Y* in a way that *Y* is owed.

To express good will to someone is to act in a way that demonstrates one's positive regard for them: we express good will when we show others that we care about them and how they fare. Further, to express good will, it is typically not sufficient to have a mere preference or background wish that they fare well, nor to merely inform them that we care about them. Instead, expressions of good will are a matter of the ways that we treat others. It is through treating others in some ways and not others that we can reveal that, over and above having a preference or wish that they fare well, their interests and welfare are sufficiently important to us that we willingly act in ways that we otherwise would not if we did not care about them and how they fare. In expressing good will to someone, we convey that we take their interests and their ends as reason-giving, or as ends of our own.

Contrast expressions of good will with expressions of ill will. In expressing ill will to someone, we need not (or need not necessarily) demonstrate that we actively care about the frustration of their ends. That would be a form of malice that need not come along with just any expression of ill will. Rather, in expressing ill will to someone, we show them that we do not care enough about their interests and ends to weigh them appropriately in our deliberation. Both good and ill will reflect the ways in which others show up in our deliberation: while good will consists in demonstrating that we take someone's interests and ends as ends of our own, ill will consists in demonstrating that we fail to give others' interests and ends sufficient weight in our deliberation.

26 McConnell disagrees, arguing that if treating others in ways that they are owed makes one a moral standout—that is, if most people violate these duties—then doing so can trigger duties of gratitude ("Gratitude, Rights, and Moral Standouts"). This is a different route to rejecting the Orthodox Thesis from the one I pursue in this paper. I want to remain neutral here on whether gratitude is obligatory with respect to moral standouts, but for plausible considerations that suggest otherwise, see Macnamara, "Gratitude, Rights, and Benefit."

There is a general connection between directed duties and ill will. Recall the claim about resentment and the demands of morality: *A* is warranted in resenting *B* only if *B* wrongs *A*. And note further the following commonly accepted Strawsonian claim about the object of resentment: resentment is (appropriately) felt toward (apparent) displays of ill will. From these two claims, it follows that part of what directed duties require of us is to refrain from acting in ways that would display ill will to others. By contrast, there is no necessary connection between directed duties and good will. Treating others in ways that they are owed need not thereby display good will—indeed, it need not display any quality of will whatsoever.

However, there is a specific class of duties that does have a necessary connection to good will. These are duties that not only require us to avoid acting in ways that display ill will to others but, further, require us to act in ways that display good will to others. And I think that the duties involved in the four cases in section 1 are members of this class—in other words, they are duties of good will. Why think that these duties require *X* to act in ways that display good will rather than merely requiring *X* to avoid acting in ways that display ill will?[27]

Start with Supermarket: *X* owes it to *Y* to help by picking up the cans. Does doing so convey that *X* takes *Y*'s interests and ends as ends of *X*'s own? The answer seems to be yes, at least in a limited way. In helping by picking up the cans, *X* does not demonstrate that *X* takes *all* of *Y*'s ends as ends of their own, just in virtue of these ends being *Y*'s ends. But *X* does demonstrate that they take a particular end of *Y*'s as an end of their own—namely, *Y*'s end of bringing the items that they had selected to the cashier. *X* does not (unless the case is further specified in strange ways) have as an independent end of their own that *Y* brings the items that *Y* had selected to the cashier. Rather, *X* adopts this end because it is *Y*'s end and because *X* notices *Y* in need of help in achieving this end.[28] In requiring *X* to help by picking up the cans, then, *X*'s duty of (minor) aid or beneficence requires *X* to act in a way that expresses good will to *Y*.

For similar reasons, *X*'s duty of rescue in Beach Rescue requires *X* to act in a way that conveys good will to *Y*. In jumping into the water and attempting to

27 In arguing that the duties involved in the four cases are duties of good will, requiring *X* to act in ways that express good will to *Y*, I do not mean to claim that any of these duties are always duties of good will. For instance, I do not mean that all duties of beneficence require agents to express good will in the sense described. I mean only that the specific duties that *X* is subject to in these cases are duties of good will.

28 Note that it does not follow that *X* would not convey ill will in refraining from helping; rather, *X*'s choice situation involves choosing between an option that would express ill will and an option that would express good will. In situations like Supermarket, unlike others, there is no option that would be neutral with respect to the quality of will expressed in one's conduct.

save *Y* from drowning, *X* fulfills their duty of rescue. But *X* also demonstrates that they take *Y*'s interests and ends as ends of their own—*Y*'s end of staying alive, or perhaps even *Y*'s very ability to set ends at all. Now, if we specified the case differently, *X*'s lifesaving aid may not demonstrate good will—for instance, if *X* had as an independent end of their own that *Y* is saved from drowning or if *X* were coerced or otherwise pressured into helping. But *X* jumps into the water and saves *Y*'s life because *X* notices the threat to *Y*'s end of staying alive (or to *Y*'s ability to set ends at all) and adopts *Y*'s ends as ends of *X*'s own. In requiring *X* to attempt to save *Y*'s life, then, *X*'s duty of rescue requires *X* to act in a way that expresses good will to *Y*—that is, *X*'s duty of rescue is a duty of good will.

Next, in Business Competition, *X* has a duty of gratitude that requires them to refrain from opening the new store in *Y*'s area. But in refraining from opening the new store, *X* expresses good will to *Y*—*X* demonstrates that they take *Y*'s ends as ends of their own. In particular, *X* demonstrates that they adopt *Y*'s end of staying in business, and thereby protecting their livelihood, as an end of *X*'s own. Further, we can see from the case that this is not an independent end that *X* has: *X* is considering opening the new store, which would be an ordinary and (presumably) profitable business activity, and only decides not to upon learning that doing so would drive *Y*'s store out of business. So in requiring *X* to refrain from opening the new store, *X*'s duty of gratitude requires *X* to act in a way that expresses good will to *Y*.

Finally, in Hurtful Joke, *X*'s duty of apology requires *X* to sincerely apologize for hurting *Y*'s feelings. In apologizing to *Y*, *X* expresses good will to *Y*, since *X* demonstrates that *X* takes *Y*'s ends of avoiding emotional pain, and perhaps having one's friendships be mutually supportive and caring, as ends of *X*'s own. Of course, if *X* had "apologized" in other ways, *X* might not thereby express good will to *Y*—merely saying the words "I'm sorry" does not always suffice for sincerely apologizing and thus does not always fulfill a duty of apology. But given that *X*'s apology is sincere and made with an assurance of more careful sensitivity to *Y*'s emotions in the future, it does seem that *X* expresses good will to *Y*. In fulfilling the duty of apology, *X* expresses good will to *Y*, and so part of what the duty requires of *X* is to express good will.

The duties at issue in the cases presented in section 1 are thus duties of good will—part of what they require is that an agent acts in ways that express good will to another agent. I will now argue for a claim about the significance of this fact about the duties involved in the four cases: the fact that *X* fulfills a duty of good will in each case explains why *Y* owes *X* gratitude in response.

For this argument, we need not proceed case by case. Instead, we can start from a claim about what gratitude is characteristically a response to: we (appropriately) feel gratitude in response to (apparent) displays of good will. It is this

fact about the nature of gratitude that leads Strawson to describe gratitude and resentment as an opposed pair: gratitude is characteristically felt and expressed in response to displays of good will, while resentment is characteristically felt and expressed in response to displays of ill will. When *A* fulfills a duty of good will that is directed toward *B*, *A* expresses good will to *B*, and when *A* expresses good will to *B*, it is appropriate for *B* to feel and express gratitude in response. So the fact that *X* fulfills duties of good will in the four cases explains why it is appropriate for *Y* to feel and express gratitude in response.

However, even where gratitude is appropriately felt or expressed, it is not always *owed*. And I have claimed that the fact that *X* fulfills a duty of good will in each of the four cases explains not only why *Y* might appropriately feel gratitude in response but, further, why *Y owes X* gratitude in response. In order to see why *X*'s fulfillment of duties of good will explains why *Y* owes gratitude in response, it is helpful to look at a few examples of cases in which gratitude is appropriately felt or expressed but not owed. (We might think of those individuals who we would characterize as especially generous with their gratitude.) Someone might sincerely express gratitude to their boss for giving them an ordinary cost-of-living wage. Or someone might sincerely express gratitude to the organizers of a raffle upon winning the top prize. Or finally, someone might sincerely express gratitude to a pizza delivery person who delivers a pizza fairly quickly. In none of these cases does gratitude seem inappropriate or unfitting. But neither does gratitude seem owed. Gratitude is appropriate because of the benefit provided in each example, especially in virtue of gratitude's ability to maintain a happy equilibrium in the dynamics of interpersonal relationships (even quite fleeting ones, such as with the raffle organizers or the pizza delivery person).

Why is gratitude not owed in these cases? A striking fact about these cases, as opposed to the four presented in section 1, is that in none of them does the benefactor display good will to the beneficiary. The boss does not demonstrate that they take the employee's ends as ends of their own—only that they want their employees to be fairly compensated (or perhaps only that they want to retain their employees and fear that without offering such a raise, their employees will find jobs elsewhere). The raffle organizers do not demonstrate that they take the winner's ends as ends of their own—after all, supposing that it is a fair raffle, the winner is selected randomly. And the pizza delivery person, unless they are familiar with the person who ordered the pizza and accordingly makes an effort to deliver especially quickly, does not demonstrate that they take the pizza recipient's ends as ends of their own.[29] When a benefactor does not display good

29 There is another reading of these cases in which each person does display good will—but good will to the beneficiary community as a whole (the boss's employees, the raffle participants, the customers of the pizza restaurant) rather than to individuals. If that is true, then

will to their beneficiary in the provision of the benefit, it seems, the beneficiary does not owe the benefactor gratitude in response. And when a benefactor does display good will in providing a benefit, it seems, the beneficiary owes them gratitude in response. Accordingly, with respect to the four cases presented in section 1, the fact that *X* fulfills a duty of good will directed to *Y* explains not just the fact that it is appropriate for *Y* to feel and express gratitude in response but, further, the fact that *Y* owes *X* gratitude in response.

However, one might wonder whether expressing good will is really sufficient for a duty of gratitude in response or in what sense of "good will" expressions of good will trigger duties of gratitude. More specifically, some ways of treating others seem aptly described as expressing a kind of good will, but it is less than clear that gratitude is owed in response. First, we might help someone but in such a way that we do too much to take their ends as our own, leaving them too little room or opportunity to pursue their ends themselves. Call this *paternalistic good will*. This would amount to a kind of good will but at the cost of insufficient respect for them as independent agents. Second, we might help someone but purely on the basis of duty or moral rectitude instead of any concern for how they in particular fare. Call this *righteous good will*. This too would be a type of good will insofar as it involves a desire to help others (at least when required)—but seemingly not for the reasons that make gratitude called for in response. Do expressions of paternalistic and righteous good will, in combination with the provision of benefits, trigger duties of gratitude?[30]

First, it is worth getting clearer on the sense in which paternalistic good will is a type of good will. To this point, I have described good will in fairly general terms as a quality of will toward someone that involves taking their ends as ends of one's own. And paternalistic good will does seem to involve taking another person's ends as ends of one's own. Suppose that my friend is an aspiring writer, and they have asked me to proofread a short story of theirs before they submit it to literary journals, since I have published in these journals many times. I notice not only a handful of typographical mistakes but also ways in which their writing can be improved more generally. Without their knowledge, I make changes to their word choice, dialogue, and the flow of their sentences, in the hope that doing so will give them a better chance of being accepted—while still letting them maintain the belief that the work is entirely their own. Plausibly, I have

these people may be owed gratitude in response, and in particular, the relevant beneficiary communities may owe it to these people to express gratitude. Consider, as an example of this sort of communal gratitude, organizing a lunch for volunteers who clean up a neighborhood garden. This reading of these cases would only bolster my argument: it would show that when someone displays good will of the relevant kind, they are owed gratitude in response.

30 Thanks to an anonymous reviewer for raising these questions.

taken their ends as my own, since I make the changes because I want my friend to succeed in their literary endeavors. But in doing so, I rob them of the opportunity to succeed for themselves. I help them too much, and though I act with good will toward them, I do so at the cost of not treating them with proper respect.

Suppose that despite my efforts to keep my modifications a secret, my friend discovers the changes that I have made to their story. It does not seem that they would owe me gratitude for doing so—in fact, quite the opposite. My friend would be justified in feeling angry and hurt in light of my disrespectful treatment of them. Expressions of paternalistic good will, then, do not necessarily (or perhaps ever) trigger duties of gratitude. The expressions of good will that trigger duties of gratitude are expressions of *nonpaternalistic* good will. This also has consequences for how to understand duties of good will: rather than being obligations that bear no relation to respect and direct us to help others achieve their ends in whatever ways we can, duties of good will contain an implicit obligation not to help others achieve their ends in ways that involve disrespecting them in the process. Expressions of good will (in the sense in which they trigger duties of gratitude) and duties of good will (in the sense in which their fulfillment triggers duties of gratitude) should thus be understood in nonpaternalistic terms.[31]

The second question about the sense in which good will (plus the provision of a benefit) triggers duties of gratitude concerns "righteous" good will, or helping others from the motive of moral rectitude instead of concern for how a particular person fares. Imagine a variant of Supermarket in which *X* helps *Y* by picking up the cans, and in response to *Y*'s thanks, *X* tells *Y* something to the effect of "No thanks necessary—it was nothing personal, I simply aim to help others when that seems like the morally right thing to do." In some sense, *X* displays a laudable motive, as *X* is committed to treating others in accordance with duty. And further, it seems to express at least a sort of good will, since *X* takes others' ends as ends of their own, at least when morality requires that *X* do so. But because *X* acts only from rectitude and not from sincere care for *Y*, it may also seem that *X* does not display the kind of good will that calls for

31 This nonpaternalistic account of duties of good will parallels Kant's treatment of duties of virtue to others in the *Metaphysics of Morals*, 6:448. There, he argues that good will (which he calls "love") and respect can come apart in our treatment of one another, but that they are united in what duty requires of us. They are, he says, "united by the law into one duty, only in such a way that now one duty and now the other is the subject's principle, with the other joined to it as accessory." I take this to mean that the sense in which morality requires us to treat others in ways that express good will to them is limited to ways that do not involve disrespecting them, since taking someone's ends as our own in a way that involves disrespect is tantamount to using someone as a mere means to their own ends.

gratitude in response. Does *Y* still have a duty of gratitude if *X* helps *Y* out of moral rectitude instead of good will toward *Y* in particular?

In order to answer this question, it is important to distinguish between two motives that can each be characterized in terms of moral rectitude. That is, there are two meaningfully different motives that are consistent with *X*'s helping *Y* not because *X* cares about *Y* in particular but instead because of a commitment to doing what is morally required of them. On one hand, *X* might act from a commitment to moral rectitude in the sense that *X* lacks any pre-existing relationship with *Y* and accordingly lacks a commitment to helping *Y* achieve their ends independently of the situation at hand. If so, then *X* helps *Y* not because of an antecedent concern for *Y* but instead because *X* is in a position to help *Y*, and the duty of beneficence directs *X* to help by picking up the cans. On the other hand, *X* might act from a commitment to moral rectitude in the sense that *X* is motivated to help *Y* not because *X* cares about helping people but because *X* wants to be the sort of person who fulfills their moral obligations. Either way, *X* helps *Y* because *X* is committed to doing what the duty of beneficence requires of them. The motives differ with respect to *why X* cares about doing what the duty of beneficence requires of them: on one hand, *X* might care about doing so because they care about helping people and how others fare; on the other hand, *X* might care about doing so because they care strictly about fulfilling their moral obligations, independently of the effects of doing so on others.

These two motives yield different results with respect to whether *X* expresses genuine good will to *Y* in helping and, consequently, to whether *Y* owes *X* gratitude in response. The first is a type of moral rectitude insofar as *X* cares about doing what is morally required *de dicto*, but it is a type of rectitude that is consistent with expressing good will. The reason why this type of rectitude is consistent with expressing good will is that we do not need an antecedent commitment to taking a person's ends as ends of our own in order to do so in a particular situation. What the duty of beneficence requires of *X* in Supermarket is to treat *Y* in such a way that *X* expresses good will to *Y*—that is, to help *Y*, thereby taking *Y*'s ends as ends of *X*'s own. Doing so because one cares about doing what the duty of beneficence requires of one does not rule out thereby expressing good will, since we can care about doing what the duty of beneficence requires of us precisely *because* we care about helping others in general. On the other hand, though, the second type of motive does appear to be incompatible with expressing good will. If we care about doing what the duty of beneficence requires of us solely because we want to be the kind of person who does what morality requires of us, then we do not truly take others' ends as ends of our own. We treat others' ends instrumentally, as opportunities to achieve our own end of being a morally

righteous person. This amounts to a type of fetishization of the demands of morality rather than a genuine concern for others and how they fare. And to the extent that *X* helps for this reason, *X* does not express good will to *Y*, and *Y* owes *X* no gratitude in response.[32] Accordingly, either the kind of righteous good will displayed by someone who helps because of a commitment to moral rectitude is perfectly consistent with expressing good will (if their motive is of the first type) and so calls for gratitude in just the same way as being motivated by a direct concern for how someone fares, or it involves no good will toward others and so does not call for gratitude in response at all.

Let us pause to take stock of what I have argued so far. I presented four cases that I argued are counterexamples to the Orthodox Thesis, since they each have the following two features: (1) *X* treats *Y* in a way that *Y* is owed, and (2) *Y* owes *X* gratitude in response. I then argued that these cases have a further feature in common: (3) in each, *X* fulfills a duty of good will, or a duty that requires *X* to act in a way that expresses good will to *Y*. Finally, I argued that feature 3 explains why feature 2 holds in each case. We thus have not only a case against the Orthodox Thesis but also an explanation for why it is false. The Orthodox Thesis delivers the wrong verdict in cases where an agent fulfills a duty of good will. Its plausibility depends on the assumption that we are never required by duty to treat others in such a way that we express good will to them. But this assumption is false, as demonstrated by the duties at issue in the four cases.

## 4. GRATITUDE, ENTITLEMENT, AND SUPEREROGATION

I now want to consider an objection to my view based on a claim about the nature of gratitude as a feeling or emotion. This objection stems from an argument commonly given in favor of the Orthodox Thesis. The argument, roughly, is this:

*The Entitlement Argument for the Orthodox Thesis*:

1. Feeling grateful to someone involves representing what one is grateful for as something to which one is not normatively entitled.[33] (Call this the *Entitlement Claim*.)

32 Might *Y* be grateful nevertheless that *X* cares about being a morally righteous first place, rather than simply flouting the demands of morality? While it could be intelligible for *Y* to be grateful that *X* is committed to living up to the demands of morality, especially if most people *Y* interacts with regularly flout the demands of morality, it seems that gratitude *to* X *in particular* would be out of place insofar as *X* does nothing to convey good will *to* Y *in particular*.

33 I will interpret this claim to mean that feeling grateful to someone *necessarily* involves representing what one is grateful for as something to which one is not normatively entitled,

2. If the Orthodox Thesis is false, then we are sometimes morally required to be grateful for things to which we are normatively entitled.
3. It cannot be true that both (a) we are morally required to be grateful for *p*, where we are normatively entitled to *p*; and (b) we are morally required to represent *p* as something to which we are not normatively entitled.
4. Therefore, the Orthodox Thesis is true.[34]

Although offered as an independent argument in favor of the Orthodox Thesis, the Entitlement Argument can be repackaged as an objection to my view. In particular, it may seem that so long as we accept premise 3, I am committed to denying the Entitlement Claim, a premise that has intuitive appeal for many.[35] I agree with this objection that if we accept the Entitlement Claim, then my view is false. But I will argue that we can explain both why this premise is false as well as its intuitive appeal. I will first explain the effect that accepting the Entitlement Claim would have on my account of the interaction of moral duties in cases like the four presented in section 1 and will then provide an explanation of the falsity of this premise, which nevertheless vindicates its intuitive appeal.

Suppose for the moment that the Entitlement Claim is true: part of what is involved in being grateful is representing what one is grateful for as something to which one is not normatively entitled. More specifically, part of what is involved in being grateful for the way in which someone treats us is representing the way in which they treat us as something to which we are not normatively entitled. And for someone to owe it to me that they treat me in some way just is for me to be normatively entitled to them treating me in this way.[36] So gratitude

---

rather than merely *typically* involving such a representation, since the argument is invalid if premise 1 is interpreted in the latter way.

34 See Feinberg, "The Nature and Value of Rights," for an early version of this argument. See Macnamara, "Gratitude, Rights, and Benefit," for the most developed version of it, and see Attie-Picker, "Obligatory Gifts," for endorsement of the Entitlement Claim, albeit for a different purpose.

35 Premise 3 is not best justified by appeal to intuition; rather, its plausibility is better seen as stemming from something like the claim that morality cannot require us to represent the moral landscape incorrectly. I think that more would need to be said to justify this further claim—or whatever claims we might appeal to in order to justify premise 3—but for the purposes of this paper, I am happy to grant the truth of premise 3 to those who believe the Entitlement Argument to be sound.

36 I am here and throughout this section assuming that talk of what agents are "normatively entitled" to, in the context of the Entitlement Argument, is synonymous with talk of what agents are owed. But there is another sense of entitlement that we might employ: to be normatively entitled to something might mean having the ability to claim it (in Feinberg's "performative" sense of 'claim') or having the standing to demand it. If we interpret the

is out of place when others treat us in ways that we are owed—or at least we must pretend to ourselves that we were not really owed this form of treatment at all if we are to feel gratitude.

Let us start by considering Hurtful Joke. If the Entitlement Argument is sound, and if *X* does owe it to *Y* to sincerely apologize for hurting *Y*'s feelings, then *Y* cannot owe *X* gratitude in response. But it is worth looking in particular at what is entailed by the Entitlement Claim here. This claim says that one cannot feel grateful without representing what one is grateful for as something to which one is not normatively entitled. Now, suppose that in response to *X*'s apology, *Y* feels grateful and, further, expresses gratitude and forgives *X*. If the Entitlement Claim is true, then in feeling grateful for *X*'s apology, *Y* necessarily represents *X*'s apology as something that *Y* is not entitled to. But while it certainly seems possible for *Y* to represent *X*'s apology as something that *Y* is not entitled to, it hardly seems impossible for *Y* both to acknowledge that *X* genuinely did owe them an apology—to acknowledge that it would be wrong for *X* not to apologize—and also to feel grateful for *X*'s apology. The Entitlement Claim entails, counterintuitively, that unless *Y* represents *X*'s apology as something that *Y* is not entitled to, *Y* simply cannot feel grateful for the apology.

The Entitlement Claim also delivers the same verdict in Supermarket, Beach Rescue, and Business Competition. In each case, unless *Y* represents the way in which *X* treats them as something that *Y* is not entitled to, then *Y* cannot feel gratitude in response. And while it might be true that some individuals, were they in *Y*'s position, would not be disposed to represent the way in which *X* treats them as something that they are entitled to, it certainly seems possible for *Y* both to feel grateful and to acknowledge that *X* treats them in a way they are owed. Further, for additional evidence for this claim, consider Business Competition. It is possible for *Y* to either be grateful for *X*'s refraining from opening the new store (supposing that *X* refrains from doing so) or be resentful for *X* denying their request and opening the new store anyway (supposing that *X* opens the new store) without holding different beliefs about what morality requires of *X*. If *Y* resents *X* for denying the request and driving *Y* out of business, then *Y* would represent *X* as failing to treat *Y* in a way *Y* is owed—that is, *Y* would represent *X*'s refraining from opening the new store as something to which *Y* is normatively entitled. But if *X* accedes to the request, it is possible for

Entitlement Argument using this interpretation of talk of what agents are "normatively entitled" to, then my response to this objection does not have purchase. But more importantly, if the Entitlement Argument is interpreted in this way, then it no longer provides an objection to my view, since the claim that we sometimes owe gratitude in response to others treating us in ways that we are owed does not entail the claim that we have the standing to demand that they treat us in these ways, or the ability to claim such treatment.

*Y* to feel grateful to *X*. Suppose that while *X* is deciding whether to accede to *Y*'s request, *Y* knows that they will resent *X* if *X* refuses, and thereby represents *X*'s refraining from opening the new store as something to which *Y* is entitled. If *X* then accedes to the request, *Y* would not need to change their mind about what morality requires of *X* in order to feel grateful. But this is exactly what the Entitlement Claim entails.

The Entitlement Claim—the key claim in this objection to my view—thus delivers implausible verdicts about the cases presented in section 1. Nevertheless, there is something intuitively plausible about it. But this intuitive plausibility, I will now argue, stems from the resemblance between the Entitlement Claim and a nearby but importantly distinct claim about the nature of gratitude. This nearby claim is what I will call the *Good Will Claim*: feeling grateful to someone involves representing what one is grateful for as expressing good will. Like the Entitlement Claim, the Good Will Claim provides a necessary condition on the feeling or emotion of gratitude. And given a further assumption, they may even seem to be equivalent claims. I think that the intuitive plausibility of the Entitlement Claim stems from the truth of the Good Will Claim, along with acceptance of a further assumption about good will and supererogation. But I will argue that this further assumption is false, that the Entitlement Claim and the Good Will Claims are not equivalent, and that only the latter is true.

The further assumption that I have in mind is this: good will can be expressed only by supererogatory actions. While this assumption is often left implicit, it captures a commonly held view of the place of good will—and, relatedly, of gratitude—in the moral landscape.[37] What might be said in favor of this assumption? One thought is that for many duties, actions that fulfill them cannot express good will, since one can be motivated by duty rather than by good will for the individual to whom the duty is owed. This is especially plausible regarding what are sometimes called *juridical* or *perfect* duties, such as duties concerning promise, property, and bodily autonomy. But these do not exhaust the range of duties that morality provides; we are subject also to *ethical* or *imperfect* duties as well. Concerning these duties, it is often suggested that we

37 Heyd helpfully makes this assumption more explicit than most. For instance, he says that "The point of supererogatory action lies … in the good will of the agent, in his altruistic intention, in his choice to exercise generosity or to show forgiveness, to sacrifice himself or to do a little uncalled favor, rather than strictly adhering to his duty" ("Supererogation," sec. 3.3). Elsewhere, connecting this assumption to gratitude, he writes, "Gratitude is generally the mark of supererogation, for it means an acknowledgement of the gratuitous, supererogatory nature of the act for which one is grateful" ("Beyond the Call of Duty in Kant's Ethics," 319).

are only required to act in accordance with them *enough* of the time—and so acting in accordance with them on any particular occasion is supererogatory.[38] But whatever the precise sense of latitude at issue in imperfect duties, we can return to the cases presented in section 1 to see that this assumption is false. In each case, *X* treats *Y* in a way that *Y* is owed. *X*'s actions are not supererogatory but required. And yet *X*'s actions express good will to *Y*; the duties that *X* fulfills are duties of good will. Accordingly, the assumption that good will can be expressed only by supererogatory actions is false.[39]

If this assumption were true, then the Good Will Claim would entail the Entitlement Claim: to represent an action as expressing good will would be to represent it as supererogatory and thus as something to which one is not normatively entitled. But without the assumption, they are importantly different claims: it is possible to represent some action as expressing good will without representing it as something to which one is not normatively entitled. Both claims seem to aim at capturing a way in which we represent an action as freely performed and indicative of how someone really feels about us when we feel grateful for their treatment of us. But while the intuitive plausibility of the Entitlement Claim depends on an incorrect assumption about the relation between good will and the supererogatory, the Good Will Claim does not. By appealing to the Good Will Claim in tandem with the earlier discussion of the cases presented in section 1, we can explain both the intuitive appeal of the Entitlement Claim as well as its falsity. The objection to my view on the basis of the Entitlement Claim accordingly does not succeed.

We have seen that duties of good will form an important class of counterexamples to the Orthodox Thesis. Duties of good will provide cases in which one agent owes it to another to treat them in a certain way, but the second nonetheless owes the first gratitude for doing so. Others sometimes owe us treatment that expresses their good will to us. And because good will is the proper ground of gratitude, when they treat us in these ways, we owe them

38 This view would make sense of Heyd's examples in the previous footnote: generosity, forgiveness, and aid all seem to fall into the category of the ethical or imperfect. I argue elsewhere that this view faces a significant challenge in its ability to explain cases in which imperfect duties appear to require agents to perform particular actions (Segal, "The Indeterminacy of Imperfect Duties").

39 I suspect that the considerations described in this paragraph provide the bulk of the rationale for this assumption: many either think of all duties on the model of juridical or perfect duties, or else think of all actions performed in accordance with imperfect duties as supererogatory. But this is nothing more than a suspicion. Regardless, my arguments in sections 1–3 suffice to provide an independent argument for the falsity of the assumption.

gratitude in return. If this is right, then the domains of gratitude and duty are much closer than we often think.[40]

*Kansas City University*
*asegal@kansascity.edu*

## REFERENCES

Attie-Picker, Mario. "Obligatory Gifts: An Essay on Forgiveness." *Ergo* 9, no. 18 (2022): 487–508.

Berger, Fred. "Gratitude." *Ethics* 85, no. 4 (1975): 298–309.

Camenisch, Paul. "Gift and Gratitude in Ethics." *Journal of Religious Ethics* 9, no. 1 (1981): 1–34.

Darwall, Stephen. *The Second-Person Standpoint*. Harvard University Press, 2009.

Feinberg, Joel. "The Nature and Value of Rights." *Journal of Value Inquiry* 4 (1970): 243–57.

Gaus, Gerald. "The Demands of Impartiality and the Evolution of Morality." In *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*, edited by Brian Feltham and John Cottingham. Oxford University Press, 2010.

Helm, Bennett. "Gratitude and Norms: On the Social Function of Gratitude." In *The Moral Psychology of Gratitude*, edited by Robert Roberts and Daniel Telech. Rowman and Littlefield, 2019.

Herman, Barbara. "Being Helped and Being Grateful: Imperfect Duties, the Ethics of Possession, and the Unity of Morality." *Journal of Philosophy* 109, nos. 5–6 (2012): 391–411.

Heyd, David. "Beyond the Call of Duty in Kant's Ethics." *Kant-Studien* 71 (1980): 308–24.

———. "Supererogation." In *Stanford Encyclopedia of Philosophy* (Winter 2019). https://plato.stanford.edu/archives/win2019/entries/supererogation/.

Kant, Immanuel. *The Metaphysics of Morals*. In *Practical Philosophy*, translated and edited by Mary Gregor. Cambridge University Press, 1996.

Lyons, Daniel. "The Odd Debt of Gratitude." *Analysis* 29, no. 3 (1969): 92–97.

Macnamara, Colleen. "Gratitude, Rights, and Benefit." In *The Moral Psychology*

40 I owe gratitude to many people who helped me with this paper, and in particular want to thank Stephen Engstrom, Joseph Metz, Aaron Salomon, Sergio Tenenbaum, Michael Thompson, Daniel Webber, Jennifer Whiting, Karolina Wisniewska, and two anonymous reviewers for *JESP*.

*of Gratitude*, edited by Robert Roberts and Daniel Telech. Rowman and Littlefield, 2019.

Manela, Tony. "Gratitude." In *Stanford Encyclopedia of Philosophy* (Winter 2021). https://plato.stanford.edu/archives/win2021/entries/gratitude/.

———. "Obligations of Gratitude and Correlative Rights." In *Oxford Studies in Normative Ethics*, vol. 5, edited by Mark Timmons. Oxford University Press, 2015.

McConnell, Terrance. "Gratitude, Rights, and Moral Standouts." *Ethical Theory and Moral Practice* 20, no. 2 (2017): 279–93.

Segal, Aaron Eli. "Gratitude and Demand." Unpublished manuscript.

———. "The Indeterminacy of Imperfect Duties." Unpublished manuscript.

Stohr, Karen. *On Manners*. Routledge, 2012.

Strawson, P. F. "Freedom and Resentment." In *Freedom and Resentment and Other Essays*. Routledge, 1974.

Wallace, R. Jay. *Responsibility and the Moral Sentiments*. Harvard University Press, 1994.

Weiss, Roslyn. "The Moral and Social Dimensions of Gratitude." *Journal of Southern Philosophy* 23, no. 4 (1985): 491–501.

*Journal of Ethics and Social Philosophy*
VOL. 29, NO. 2 · JANUARY 2025
*https://doi.org/10.26556/jesp.v29i2.3666*


# ENCLAVES FOR THE EXCLUDED

## A PESSIMISTIC DEFENSE

*Jamie Draper*

In western liberal democracies, and especially in Europe, the politics of immigration is intertwined with the politics of integration. Approaches to integration vary across national contexts, but there are also significant points of convergence.[1] One such point of convergence is the widely held view that immigrants have a *duty to integrate* in receiving societies. In the United Kingdom, for example, successive Labour and Conservative governments have made the integration of immigrants a political priority in response to popular anxiety about immigrant communities being disconnected from the social and cultural mainstream. In October 2023, Suella Braverman, then home secretary, chastised immigrants for living "parallel lives" and "not taking part in British life."[2]

At the same time as they are expected to integrate, members of some immigrant communities are viewed and treated as inferiors in receiving societies. Anti-immigrant attitudes are widespread in Europe in general, but they are especially pronounced for some immigrant communities in particular—typically, predominantly Muslim ethnic minority communities.[3] Anti-immigrant attitudes are expressed both in media discourse and in a political culture in which immigrant minorities are stigmatized and represented as a civilizational threat.[4] And crucially, it is often precisely those immigrant communities that are most stigmatized who are the primary addressees of the demand to integrate.

This paper investigates the claim that immigrants have a duty to integrate in light of the fact that many immigrants who are expected to integrate are stigmatized in receiving societies. I argue that immigrant minorities who face a particular kind of relational inequality—social exclusion—have a moral permission

1 Joppke, "Beyond National Models."

2 Hughes, "Braverman."

3 Semyonov, Raijman, and Gorodzeisky, "The Rise of Anti-Foreigner Sentiment in Europe"; and Bell, Valenta, and Strabac, "A Comparative Analysis of Changes in Anti-Immigrant and Anti-Muslim Attitudes in Europe."

4 Brubaker, "Between Nationalism and Civilizationism"; and Saeed, "Media, Racism and Islamophobia."

to form enclaves. Enclaves, as I understand them here, conflict with at least some putative duties to integrate. So my argument suggests that immigrants who face social exclusion have, at most, limited duties to integrate.

My defense of enclaves for the excluded involves a positive argument and a negative argument. Positively, I argue that enclaves can play an important role in supporting the self-respect of members of socially excluded groups. Social exclusion is a threat to self-respect, and enclaves can have a protective function for those whose self-respect is threatened in this way. Negatively, I argue social exclusion makes the duty to integrate unreasonably burdensome. I also argue that even if integration is a genuine duty, it cannot be permissibly enforced as a social expectation vis-à-vis socially excluded immigrants, because members of dominant social groups lack the standing to blame socially excluded immigrants for failing to integrate.

But while I argue that socially excluded immigrants have only limited duties to integrate, I also accept that integration can be an important way of combatting relational inequality. My argument thus has a pessimistic conclusion: social exclusion means that immigrant minorities have at best only limited duties to integrate, but it is in the context of social exclusion that integration is particularly valuable.

My focus in this paper is on the integration of *immigrants* in particular, and I focus on first-generation, voluntary immigrants. To the extent that they face both social exclusion and the demand for integration, however, my argument also applies to second- and third-generation immigrants. It may also extend to other, nonimmigration contexts in which minorities face both social exclusion and the demand for integration, such as racial segregation in the United States, although there are clearly significant differences between these contexts. But the primary context that motivates my inquiry is that which Sune Lægaard calls "euro-multiculturalism," in which it is immigrant communities—typically ethnic and religious minorities—who are the primary addressees of demands to integrate.[5] And as we will see, there is an objection to my argument that applies to first-generation, voluntary immigrants in particular: that those who have migrated voluntarily have waived their moral permission to form enclaves. Voluntary immigrants thus represent a hard case for my argument. If I can show that voluntary immigrants have a moral permission to form enclaves when they face social exclusion, then this bears well on the prospects for my argument more generally.

5 Lægaard, "Unequal Recognition, Misrecognition and Injustice." See also Holtug, *The Politics of Social Cohesion*, 23–37.

The paper proceeds as follows. First, in section 1, I clarify three central concepts involved in my argument: integration, enclaves, and social exclusion. Then, I make the positive and negative arguments for my central claim: the positive argument from self-respect (section 2.1) and the negative argument from unreasonable burdens and standing (section 2.2). I then consider two objections to my argument: that those who have migrated voluntarily have waived their moral permission to engage in enclave formation (section 3.1) and that enclaves may hinder the pursuit of relational equality (section 3.2). Finally, in section 4, I conclude by highlighting a virtue of my argument and an upshot of my argument for debates about immigrant integration.

## 1. INTEGRATION, ENCLAVES, AND SOCIAL EXCLUSION

'Integration' can refer both to a *state* and to a *process*. A state of integration exists when there are no significant patterns of differentiation between members of different social groups. Conversely, a society is segregated to the extent that its members are differentiated according to their membership in different social groups. We can imagine a continuum with a fully integrated society at one end a fully segregated society at the other, with a society being more or less integrated according to its degree of differentiation by social group membership.

As a process, integration refers to a dynamic of *mutual adjustment* between majorities and minorities that brings a society into a more integrated state. This process of mutual adjustment may involve changing norms and expectations, patterns of behavior and social practices, and/or values and beliefs. This dynamic of mutual adjustment is what distinguishes integration from *assimilation*, where one group—typically a minority group—adjusts to the norms, values, customs, and behaviors of another.[6] Integration can also vary along different dimensions. David Miller distinguishes between *civic*, *cultural*, and *social* forms of integration.[7] My focus in this paper is primarily on social integration, although these three forms of integration are often intertwined in practice. Social integration refers to people regularly interacting with each other in a range of social contexts, for example by working alongside each other, living in the same neighborhoods, attending the same schools, joining the same associations, and mixing socially in friendships and marriages.[8]

6 Parekh, *Rethinking Multiculturalism*, 219–24; Modood, *Multiculturalism*, 48; Mason, "The Critique of Multiculturalism in Britain"; and Klarenbeek, "Reconceptualising 'Integration as a Two-Way Process.'"

7 Miller, *Strangers in Our Midst*, 132–33.

8 Miller, *Strangers in Our Midst*, 132. See also Anderson, *The Imperative of Integration*, 116–17.

This account of integration is not moralized. Nothing in the concept of integration itself—either as a state or a process—means that it is morally valuable. But advocates of integration have argued that it is a morally valuable social goal, for example because it sustains support for just institutions, promotes social cohesion, or creates a shared national identity.[9] Miller argues that social integration is valuable for two reasons. First, it is valuable because it opens up greater opportunities for immigrants themselves. And second, it is valuable because it creates a basis of trust and helps to prevent intergroup conflict in society more broadly.[10]

If integration is a valuable social goal, then immigrants may have a moral duty to participate in the process of integration. What exactly would such a "duty to integrate" involve? If we focus on social integration, then this duty would involve immigrants orienting their patterns of social interaction towards the receiving society as a whole rather than only or predominantly towards other members of a community to which they belong. This could involve, for example, mixing socially with nonimmigrants in friendships, workplaces, and voluntary associations. If there is a duty to integrate in this sense, then it is an *imperfect* duty to engage in these behaviors to a sufficient degree rather than a perfect duty to interact with any particular individual, and what counts as sufficient will vary on different views of the duty to integrate. Given the two-way nature of the process of integration, the duty to integrate could also be a *conditional* duty, which would mean that immigrants have a duty to integrate only if nonimmigrants also do their part in the process of integration. As we will see, my argument does suggest that this is a fruitful way of understanding the duty of integration. But this is something to be established through argument rather than something to be built into the idea of the duty to integrate from the start. For the moment, we can treat the duty to integrate as a putative duty owed by immigrants to mix socially to a sufficient degree with nonimmigrants within the receiving society.

Depending on which justification we give for integration, the duty to integrate may be owed either to immigrants themselves or to the receiving society more broadly. For the moment, I will treat the duty to integrate as a duty owed to the receiving society. I do so because this is how the duty to integrate is often implicitly understood when it is invoked in public claims that immigrants ought to integrate. In the later part of the paper, however, I return to the idea

9 See, respectively, Mason, *Living Together as Equals*; Holtug, *The Politics of Social Cohesion*; and Miller, *Strangers in Our Midst*.

10 Miller, *Strangers in Our Midst*, 133–34.

that the duty to integrate might be justified by reference to the interests of immigrants themselves.

Joseph Carens distinguishes three ways that we might think about the duty to integrate: as a *requirement*, as an *expectation*, and as an *aspiration*.[11] A requirement is a duty that is explicit and legally enforceable, such as a citizenship test or an integration class with penalties for nonparticipation. An expectation is an informal duty that is enforced through social sanctions, such as a social norm according to which those who do not integrate are liable to blame. An aspiration is a mere hope that immigrants will integrate, without any formal or informal sanctions attached. My focus is on the *expectation* that immigrants should integrate, as well as the corresponding social sanctions that are applied to immigrants who do not.[12] This focus allows us to examine behaviors that are not usually enforced through legal restrictions, such as mixing socially in friendships, voluntary associations, and workplaces.

The second concept that plays a central role in my argument is the concept of the *enclave*. An enclave is a pattern of social differentiation in which members of a minority social group cluster together, spatially and/or socially, in ways that they can reflectively endorse. Members of a social group who form enclaves typically see themselves as deriving some benefits—some of which I explore below—from clustering together.

This account of enclaves is somewhat broader than the way that the term is sometimes used in the social sciences. In urban geography, the concept of the enclave is used to refer specifically to a pattern of *spatial* differentiation. Peter Marcuse characterizes an enclave as "a spatially concentrated area in which members of a particular population group, self-defined by ethnicity or religion or otherwise, congregate as a means of enhancing their economic, social, political and/or cultural development."[13] Here, I use the term more broadly to refer to a pattern in which members of a social group cluster together in social and/or spatial terms. In this broader sense, members of a social group may form an enclave even if they are not spatially clustered, if their patterns of social interaction differentiate them from others in the broader society. But typically, social and spatial enclave formation will go hand in hand, since patterns of social interaction and patterns of residence are closely connected.

11 Carens, "The Integration of Immigrants," 30–31.

12 My focus is on the *normative* expectation that immigrants should integrate, which is accompanied by blame when it is not fulfilled, rather than on the *descriptive* expectation that immigrants will integrate, the nonfulfillment of which might generate other reactions such as surprise or confusion. For this distinction, see Bicchieri, *The Grammar of Society*, 13–15.

13 Marcuse, "The Enclave, the Ghetto and the Citadel," 242.

Marcuse also distinguishes enclaves from another pattern of social differentiation: *ghettos*. A ghetto is "a spatially concentrated area used to separate and to limit a particular involuntarily defined population (usually by race) held to be, and treated as, inferior by the dominant society."[14] A ghetto differs from an enclave in that it is an involuntary form of segregation that is imposed on a social group. Enclaves, by contrast, are usually understood as involving at least some degree of *self-segregation*.[15] Understood in this way, enclaves and ghettos are ideal types. In reality, the lines between them are blurred. In many of the neighborhoods that social scientists characterize as enclaves, patterns of social differentiation are likely to result from a mixture of involuntary constraints, such as limited availability of affordable housing, and voluntary decisions, such as a desire to live in a neighborhood with others who speak the same language or have similar customs and lifestyles.[16] For this reason, I think it is better to say that enclaves are patterns of social differentiation that can be *reflectively endorsed* by their members rather than to say that they are the consequence of perfectly voluntary decisions. This conception allows for a pattern of social differentiation to be an enclave even if there is some degree of involuntariness in its causal genesis, if its members nonetheless affirm their participation within it, or would do so upon reflection. One test for whether a pattern of social differentiation is an enclave is whether those who participate in it do so even though there are real opportunities for them to do otherwise, for example by changing jobs, moving house, or participating in different cultural or social activities. These alternative opportunities need not be cost-free, but they should be effectively open to immigrant minorities.

Enclaves conflict with integration in the sense that the more enclaves there are in a society and the more pronounced those enclaves are, the less that society is in an integrated state. But enclaves need not involve *total* separation from other social groups. The existence of enclaves is compatible with some degree of integration in a society. Integration is a matter of degree, and the degree to which a society is integrated will depend in part on how pronounced and widespread enclaves are within it.

The third concept that plays a central role in my argument is *social exclusion*. Social exclusion, as I understand it here, is a particular kind of relational inequality. Social hierarchies are durable, systematic inequalities between members of different social groups that are sustained by norms, rules, and

14 Marcuse, "The Enclave, the Ghetto and the Citadel," 231.

15 Soja, *Seeking Spatial Justice*, 54–56.

16 De Haas, *How Migration Really Works*, 182–95; and Portes and Manning, "The Immigrant Enclave."

habits.[17] Elizabeth Anderson distinguishes between hierarchies of *command*, *standing*, and *esteem*.[18] Hierarchies of command involve asymmetric relationships of command and obedience in which those in superior social positions hold unaccountable and arbitrary power over those in inferior social positions. Hierarchies of standing involve practices and institutions whereby the interests of those in superior social positions are given greater weight than the interests of those in inferior social positions. And hierarchies of esteem involve the stigmatization of those in inferior social positions and the valorization of those in superior social positions.

Social exclusion is a hierarchy of esteem in this sense (or a "disparity of regard," in Niko Kolodny's terms).[19] Those who are socially excluded are placed in inferior social positions in a pervasive hierarchy of esteem. They are "subject to publicly authoritative stereotypes that represent them as proper objects of dishonor, contempt, disgust, fear, or hatred on the basis of their group identities."[20] Social exclusion may involve explicit prejudice or hate speech, but most often it involves less explicit forms of prejudice that pervade informal and quotidian interactions between members of different social groups and are legitimated by socially influential ideologies.[21] Following Cécile Laborde's use of the term in her discussion of the treatment of Muslim minorities in France, I use 'social exclusion' to refer to these hierarchies of esteem.[22]

Many immigrants are socially excluded in this sense. Anti-immigrant prejudice is widespread in Europe, and it typically intersects with racial and ethnic prejudice.[23] Many of those immigrants who face the most public pressure to participate in integration—in Europe, they are typically members of Muslim minority groups—face this kind of social exclusion. There are competing explanations for these patterns of anti-immigrant prejudice in the sociological literature, with some studies emphasizing the role of perceived competitive threat, some the role of a perceived clash of values, and some the role of racial

17 Anderson, "Equality," 42.

18 Anderson, "Equality," 42–44; and Kolodny, *The Pecking Order*, 91–95.

19 Kolodny, *The Pecking Order*, 103–16.

20 Anderson, "Equality," 43.

21 See McTernan, "Microaggressions, Equality, and Social Practices"; and Haslanger, *Resisting Reality*, 446–78.

22 Laborde, *Critical Republicanism*, 202–28.

23 Bell, Valenta, and Strabac, "A Comparative Analysis of Changes in Anti-Immigrant and Anti-Muslim Attitudes in Europe"; and Semyonov, Raijman, and Gorodzeisky, "The Rise of Antiforeigner Sentiment in Europe."

biases.[24] For our purposes, the ultimate source of anti-immigrant prejudice is less important than its effect in creating a pervasive hierarchy of esteem that puts members of some immigrant groups in inferior social positions.

## 2. ENCLAVES FOR THE EXCLUDED

My central claim in this paper is that members of immigrant groups that face social exclusion have a moral permission to form enclaves. Correspondingly, they have no moral duty to participate in forms in integration that would be inconsistent with their forming enclaves. But before defending this claim, it is worth pausing to reflect on whether and why enclaves for the excluded stand in need of defense.

First, it is worth pointing out that many immigrants *do* want to integrate, and many do so wholeheartedly. In fact, the main challenge in this area is typically that immigrants who want to integrate face significant barriers, such as discrimination in labor and housing markets.[25] But the fact that many immigrants do in fact integrate does not make the question of whether they have a duty to do so irrelevant. For one thing, a defense of enclaves recasts the moral significance of the integration of socially excluded immigrants. If successful, it shows that many socially excluded immigrants do in fact integrate, despite the significant barriers that they face, *even though they have no duty to do so*. For another, some immigrants choose *not* to participate in integration, and others might do so only because of the social expectation that they ought to do so. In order to evaluate these choices, we need to know whether this expectation can be morally justified.

Second, we might think that members of socially excluded immigrant groups—like everyone else—are simply entitled to decide for themselves with whom they want to associate. Freedom of association, at least as it is conventionally understood, entitles members of immigrant groups to make decisions for themselves about with whom they want to associate, without interference by the state.[26] If we are committed to freedom of association, we might think that enclaves do not stand in need of justification in the first place.

Even if we are committed to freedom of association, we still have two good reasons to investigate whether socially excluded immigrants have a moral

24 See, respectively, Quillian, "Prejudice as a Response to Perceived Group Threat"; Schneider, "Anti-Immigrant Attitudes in Europe"; and Gorodzeisky and Semyonov, "Not Only Competitive Threat but Also Racial Prejudice."

25 De Haas, Castles, and Miller, *The Age of Migration*, 297–316.

26 For a critical evaluation of this conventional understanding of freedom of association, see Brownlee, "Freedom of Association."

permission to form enclaves. First, we can evaluate the associative choices that people make even if we do not believe that the state should intervene in those choices. It is important to understand whether and why immigrants have a moral duty to integrate that can be enforced as a *social expectation* in the receiving society, even if this has no implications in terms of the state's actions. Second, states' policy choices inevitably shape the social environment—and thereby patterns of intergroup interaction—even without directly interfering with anyone's associative choices. Housing and planning policy, subsidies and exemptions for different kinds of associations and activities—these make some associative choices easier or more attractive than others, even if that is not their primary goal. Although my argument does not by itself imply that the state has any positive duty to *facilitate* enclave formation, it does articulate reasons that suggest that the state ought not to use these tools to *undermine* enclaves.

### *2.1. The Positive Argument: Self-Respect*

The positive argument for the permissibility of enclave formation says that a moral permission to form enclaves is justified because enclaves can play an important role for immigrants in protecting their self-respect in the face of social exclusion.[27] The basic idea here is that social exclusion is a threat to self-respect, and enclaves can be an important way for socially excluded immigrants to maintain their self-respect. So what exactly is self-respect, how does social exclusion threaten it, and how can enclaves protect it?

Self-respect is, in general terms, a "sure confidence in the sense of one's own worth."[28] Philosophers tend to divide self-respect into two subtypes. The first type, *appraisal self-respect*, or *standards self-respect*, is a merit-based form of self-respect.[29] Appraisal or standards self-respect is about living up to certain (moral, practical, aesthetic) standards associated with one's self-conception in terms of life plans and projects. A musician might have appraisal or standards self-respect when they live up to the standards associated with their self-conception as a

27 Michael Merry has drawn on the concept of self-respect to defend what he calls "voluntary separation" for some minority groups. See Merry, *Equality, Citizenship, and Segregation* and "Equality, Self-Respect and Voluntary Separation." My argument differs from Merry's in two ways. First, Merry is focused on schooling, which raises some distinct concerns, whereas my argument is directed at social interaction more broadly. Second, Merry's argument from self-respect primarily aims to show that integration on unequal terms undermines self-respect. My argument develops an explanation of why social exclusion amounts to a threat to self-respect and how enclaves can protect against that threat, which draws on recent developments in the literature on self-respect.

28 Rawls, *A Theory of Justice*, 38.

29 Darwall, "Two Kinds of Respect," 39; and Schemmel, "Real Self-Respect and Its Social Bases," 631–32.

musician by practicing the violin every day. The second type, which I am primarily interested in here, is often called *recognition self-respect* or *standing self-respect*.[30] This is a non-merit-based form of self-respect that is about one's own assessment of one's status in relationships with other people. Recognition or standing self-respect—or just *self-respect*, as I will refer to it—involves the conviction that one is a moral equal of others and that one is entitled to be treated in a way that is commensurate with one's moral equality. Thus understood, self-respect plays an important role in our lives as practical agents. On the Rawlsian view, self-respect gives us justified confidence in our two "moral powers"—the sense of justice and the capacity to develop and carry out a conception of the good—and plays a role in stabilizing just institutions.[31] But more broadly, self-respect's importance lies in its role in orienting our practical commitments as agents by enabling us to see ourselves as moral equals to others.

Self-respect is not the sort of thing that we can distribute directly. But we can arrange our social and political institutions in ways that are conducive to people developing a sense of self-respect. In doing so, we can distribute the "social bases of self-respect": the features of our societies that make us secure in our conviction of our own moral worth.[32] When our social institutions put us in a position where we can be secure in our sense of our own worth as moral equals, then they have secured for us the social bases of self-respect.

As this suggests, self-respect is partly a matter of an agent's own evaluation of their moral status and partly a matter of the social conditions that enable agents to make judgments about their moral equality. One central part of the social conditions relevant to self-respect is the treatment that we receive from others. Rawls suggests that self-respect "normally depends upon the respect of others."[33] Because self-respect is partly about one's status vis-à-vis others in a society, the respect (or disrespect) that we receive from others has an important bearing on how we view ourselves as moral agents. Our relationships with others are important points in anchoring our practical self-understanding as moral agents. This idea has found expression in theories of recognition, such as Axel Honneth's analysis of self-respect as being developed through an intersubjective process of mutual recognition and vulnerable to damage through misrecognition.[34]

30 Darwall, "Two Kinds of Respect," 38; and Schemmel, "Real Self-Respect and Its Social Bases," 631–32.

31 See Krishnamurthy, "Completing Rawls's Arguments for Equal Political Liberty and Its Fair Value."

32 Rawls, *A Theory of Justice*, 54.

33 Rawls, *A Theory of Justice*, 155.

34 Honneth, *The Struggle for Recognition*. See also Margalit, *The Decent Society*.

This aspect of self-respect explains why social exclusion—being at the bottom end of a pervasive hierarchy of esteem—amounts to a threat to self-respect. Social exclusion is a signal that others do not consider you or members of your social group to have standing as their moral equal. When others treat you as their social inferior, they communicate that your "respect-standing" is lower than theirs, perhaps by behaving as if you are an object of pity or disgust or as if they have the right to expect deference and servility from you.[35] This can have an important bearing on one's self-respect. Those who are conscious of the way that others view them (and members of their social group more broadly) may internalize these views and thereby come to view themselves in terms of the prejudicial and stereotyping attitudes and norms that shape their social environment. As Emily McTernan puts it, those who face disrespect from others "lack the sort of respect from others required to underpin status self-respect, in lacking the status or standing within society that is required for it."[36] In this way, they are at risk of losing their sense of self-respect, or at least having it damaged or shaken.

Socially excluded immigrant minorities are confronted with this kind of threat to their self-respect. When they are represented by stereotypes as inferior to others, treated as such in everyday interactions with others, and denigrated by hate speech, they may come to lose their secure conviction of their own worth as moral equals of others.[37] Stereotypes that represent immigrant minorities as the proper object of fear, disgust, and contempt may be dominant among the cultural scripts and social resources that are available for agents to draw on in developing their self-conception. Socially excluded immigrant minorities may come to view themselves through the eyes of others who denigrate and mistreat them in their social interactions, and consequently they may lose a secure belief in their own moral worth. Of course, it is by no means the case that all socially excluded immigrant minorities comprehensively lose their self-respect in the face of social exclusion. But social exclusion is at least a *threat* to the self-respect of immigrant minorities.

One important challenge to this account of the relationship between social exclusion and self-respect comes from Colin Bird, who argues that a lack of respect from others does not constitute a *good reason* to lose confidence in one's own moral worth.[38] Responding to this challenge is important, not only because the challenge constitutes an objection to the argument from

35 Wolff, "Fairness, Respect, and the Egalitarian Ethos," 107.

36 McTernan, "The Inegalitarian Ethos," 95.

37 Seglow, "Hate Speech, Dignity and Self-Respect."

38 Bird, "Self-Respect and the Respect of Others."

self-respect but also because responding to it helps to illuminate why enclaves in particular can protect against the threat to self-respect posed by social exclusion. On Bird's view, losing one's self-respect in response to disrespectful treatment by others is simply not an appropriate reaction: how others treat you should have no bearing on how you view yourself. If a person loses their conviction of their own moral worth in response to mistreatment by others, then they never had any self-respect in the first place. Those with self-respect are able to withstand disrespectful treatment by others because of their disposition to view their own moral worth as inviolable, not as something that depends on the judgments of others.

This objection raises an important point: self-respect must be to at least some degree *robust*.[39] If self-respect is to play the role that it is supposed to play in orienting our lives, then it needs to be something that we can maintain in the face of at least some adversity. If self-respect were so fragile that it crumbled at the first sign of challenge, then it is not clear that it could play this role.[40] As Christian Schemmel puts it, "trying to protect people against all conceivable threats to their self-respect would mean, in effect, to try to relieve them of the need to have any."[41] But as Schemmel argues, this does not mean that we need to adopt the stoic view that self-respect has no social bases. The constitution of our practical identities is clearly at least partly social, and so the stoic view has an implausible view of the development of self-respect.[42] Given the social nature of self-evaluation, people are understandably and inevitably influenced by the treatment they receive from others in their evaluation of their own worth. Their self-evaluation can be affected by the social and cultural scripts that predominate in their social environments, which provide lenses through which they can interpret their own moral status. But with the right resources, people can retain a sense of self-respect even in the face of threats to it. Schemmel argues that the social bases of self-respect consist of the "motivational and epistemic resources to arrive at, and retain, correct convictions of [one's] own worth, even under injustice."[43]

My suggestion is that enclaves can serve the function of enabling agents to maintain their self-respect under conditions of social exclusion. Enclaves have features that make them well suited to enabling agents to respond to the threat to their self-respect posed by social exclusion. In the empirical literature on

39 Schemmel, "Real Self-Respect and Its Social Bases."

40 Schemmel, "Real Self-Respect and Its Social Bases," 637.

41 Schemmel, "Real Self-Respect and Its Social Bases," 633.

42 Bratu, "Self-Respect and the Disrespect of Others."

43 Schemmel, "Real Self-Respect and Its Social Bases," 633.

immigrant enclaves, it has been suggested that enclaves can serve as "sources of mutual support" for those who face discrimination within society.[44] We can reconstruct this idea in terms of two main ways in which enclaves can have this protective function.

First, enclaves can enable those who face social exclusion to maintain their self-respect by shaping their social environment in a way that makes stigmatizing attitudes, judgments, and stereotypes less salient in comparison to alternative cultural scripts and social resources. By orienting their social lives towards other members of their social group, socially excluded immigrants can limit their exposure to stereotypes and stigmatizing attitudes of dominant majorities and increase their exposure to the attitudes of other members of their social group. To the extent that other members of one's social group are likely to affirm more positive attitudes, to disrupt stereotypes, or to recast negatively valanced claims in more positive terms, enclave formation can thus enable socially excluded immigrants to reshape their social environment in ways that are conducive to the development of self-respect. When socially excluded immigrants are more exposed to positive representations of their own social group, they may be less influenced by pervasive stereotypes and more inclined to see them as mistakes. They may draw on alternative sets of cultural scripts in developing their self-conceptions and so may be less likely to internalize attitudes and views that cast them as inferior to others. One way that this can manifest is in enjoying a sense of belonging to a social or cultural community that combats a sense of exclusion from the dominant majority. In this way, enclaves can enable members of socially excluded immigrant groups to limit the influence that social exclusion has on the development of their self-respect.

Second, enclaves can help socially excluded minorities to develop the epistemic and motivational capacities to resist their own social exclusion. The idea that resistance to injustice can help the oppressed to maintain their self-respect is widespread in the literature on self-respect.[45] Resisting one's oppression is a way of affirming one's moral worth in the face of assaults to it. My suggestion is that enclaves can function as an epistemic and motivational resource that can enable resistance to social exclusion.

As an epistemic resource, enclaves enable those who face social exclusion to come together, share experiences, and develop common interpretive frameworks for understanding their own situations.[46] The importance of these kinds

44 Portes and Manning, "The Immigrant Enclave," 48.

45 See, for example, Boxill, "Self-Respect and Protest"; and Hay, "The Obligation to Resist Oppression."

46 Young, *Inclusion and Democracy*, 81–120. See also Draper, "Gentrification and Everyday Democracy."

of discursive spaces for the development of a critical consciousness among the oppressed has been stressed by both democratic theorists and standpoint epistemologists, who typically stress that an epistemically privileged standpoint of the oppressed is something that is achieved rather than given.[47] Coming together in enclaves can enable members of socially excluded immigrant groups to develop the epistemic and hermeneutical resources that they need to understand and contest their social exclusion.

As a motivational resource, enclaves enable members of a social group to develop the ties of intragroup solidarity that play an important motivational role in resisting social exclusion. Solidaristic relationships involve mutual identification as members of a group and a disposition to act together in pursuit of a shared goal, such as overcoming injustice.[48] Mutual identification and the disposition to act together make solidaristic relationships motivationally efficacious: they enable group members to solve coordination problems and to trust each other to do their part in collective action. Enclaves can help to build the dispositions and attitudes involved in solidaristic relationships. Those who socialize together in enclaves are more likely to mutually identify with each other and to view each other as trustworthy cooperators in shared projects, including the project of resisting their own social exclusion. Indeed, these features of enclaves have been identified by scholars of social movements as important in translating general sociological attributes like race, class, and immigration status into meaningful political identities that enable collective action.[49]

Enclaves can thus enable socially excluded immigrants to develop the epistemic and motivational resources that they need to resist their own social exclusion and, in so doing, to maintain their self-respect. The idea is not that by forming enclaves, immigrant minorities are able to eliminate injustice and oppression by engaging in resistance. It is rather that engaging in resistance—regardless of its ultimate success—is a way of affirming one's moral worth in the face of assaults against it. Since enclaves can enable resistance to injustice, they can enable the socially excluded to affirm their moral worth and so to protect their self-respect in conditions of adversity.

Of course, there are limits to these protective functions of enclaves. For one thing, there is no guarantee that the social attitudes expressed by other members of one's own social group always affirm rather than denigrate. In some contexts,

47 For democratic theory, see, for example, Mansbridge, "Everyday Talk in the Deliberative System"; Bohman, *Public Deliberation*, 132–42; and Fraser, "Rethinking the Public Sphere." For standpoint epistemology, see Toole, "Recent Work in Standpoint Epistemology"; Medina, *The Epistemology of Resistance*; and Fricker, *Epistemic Injustice*.

48 Sangiovanni and Viehoff, "Solidarity in Social and Political Philosophy."

49 Nicholls, "Place, Networks, Space"; and Castells, *The City and the Grassroots*.

stigmatizing attitudes may have become widely internalized. For another thing, the social attitudes expressed by other members of one's social group may themselves be confining. When the social identity categories that are salient in enclaves are tightly scripted, they may present an overly restrictive conception of what it means to be a member of the social group.[50] Tightly scripted social identity categories may even involve harmful social norms that can themselves undermine the self-respect of those who do not conform to intragroup norms about what kinds of behaviors or beliefs are expected of group members.

These potential costs to enclave formation are important, but they are not a reason to reject the idea that socially excluded immigrants have a moral permission to form enclaves. The costs and benefits of enclave formation will vary from person to person across different contexts, depending on—among other things—how vulnerable a person's self-respect is to the threat posed by social exclusion, how loosely or tightly scripted the identity categories in a particular enclave are, how much a person identifies with the social identity that is fostered within an enclave, and so on. Enclaves are one tool for protecting self-respect, and they are, for some, a valuable way of protecting against the threats posed by social exclusion. But for others, the costs of participating in enclaves may be too high relative to their benefits in terms of self-respect. This is why my claim is that socially excluded immigrants have a *permission* to form enclaves rather than a duty to do so. Those for whom the costs of enclave participation are too high are entitled to real opportunities to participate in other, more integrated forms of association. And where enclaves involve harmful intragroup social norms, we can object to the content of those norms without objecting to the idea that members of socially excluded groups have a moral permission to form enclaves.

These two features of enclaves explain why socially excluded immigrants have a moral permission to form enclaves. Forming enclaves can be an effective way to mitigate the threats to their self-respect posed by social exclusion, either by reducing the influence that social exclusion has in the development of one's self-conception or by enabling socially excluded immigrants to develop the epistemic and motivational resources to resist their social exclusion and thereby to reaffirm their moral worth.

### *2.2. The Negative Argument: Unreasonable Burdens and Standing to Blame*

The negative argument for the permissibility of enclave formation rejects the claim that socially excluded immigrants have a moral duty to integrate, at least in ways that would be inconsistent with their forming enclaves. I argue both

50 Darby and Martinez, "Making Identities Safe for Democracy."

that the duty to participate in social integration is unreasonably burdensome when imposed upon socially excluded immigrant minorities and that even if socially excluded immigrants do have a duty to integrate, this duty cannot be enforced as a social expectation because dominant majority groups lack the standing to blame socially excluded immigrants for failures to integrate.

The first part of this argument is that the duty to integrate is unreasonably burdensome when it is imposed upon those who face social exclusion. Immigrants who face social exclusion are put in inferior social positions in the pervasive hierarchy of esteem, and so they can reasonably expect to be treated as inferior by members of dominant or majority social groups. In modifying their patterns of social interaction to orient their social lives more towards members of the majority social group, socially excluded immigrant minorities can expect to confront stigma and hostility. They may, for example, feel pressure to modify their behavior or appearance in order to avoid aversive reactions on the part of members of the majority social group. There are costs that are imposed on socially excluded immigrant minorities when they are required to integrate socially, and my claim is that it is unreasonable to require socially excluded immigrants to bear such costs.[51]

The point here is not that it is unjustifiable to impose *any* costs on immigrant minorities in order to achieve the social goal of integration. If integration is a valuable social goal, then everyone—including immigrant minorities—may have a duty to bear some costs in order to achieve it. The integration of those from different backgrounds with different expectations and cultural practices is a morally fraught process, even in the absence of any injustices. It requires mutual accommodation and the development of "multicultural manners," whereby different parties learn to give way at some points.[52] All of this might involve immigrant minorities bearing some costs in the process of integration.

But even if it is reasonable to expect immigrants to bear some burdens in the process of integration, the duty to integrate may still be unreasonably burdensome when it is imposed upon those who face social exclusion. There are two possible interpretations of the claim that it is unreasonable to require socially excluded immigrants to bear the burdens associated with integration. The first is simply that the burdens associated with integration may be too high in the context of social exclusion. Being exposed to stigma and hostility in everyday social interactions—or even having to live with the expectation that one

51 This claim is parallel to an argument made by Tommie Shelby that Black Americans living in segregated neighborhoods have no duty to participate in integration because requiring them to participate would impose unreasonable burdens upon them (*Dark Ghettos*, 73–76).

52 Levy, "Multicultural Manners."

*might* be exposed to stigma and hostility in every interaction—is a real cost that might be unjustifiable to impose on immigrant minorities in the name of integration. This is a claim about the total burdens that can be justifiably imposed upon immigrant minorities. On this interpretation, the central claim of the negative argument is that the burdens that socially excluded immigrants face in the process of integration are simply too high for the duty to integrate to be justified. This claim will be plausible in many contexts, especially where stigma and hostility are widespread. But its overall plausibility may depend on particular features of the context in which integration takes place, which may affect precisely how costly integration is for socially excluded immigrants.

The second interpretation of this claim is that it is unreasonable to expect socially excluded immigrants to bear the burdens of integration, given that dominant majority groups are creating those costs by failing to do their part in the process of integration. As we have seen, in contrast to assimilation, integration is typically understood as a two-way process, where both minorities and majorities mutually adjust their behaviors, values, or practices. On this picture, integration may well involve some costs for both majorities and minorities, but these costs are shared and represent a fair compromise that requires that each do their part in the process of integration. But in the context of social exclusion, dominant majority groups do not hold up their end of the bargain: sustaining a pervasive hierarchy of esteem that puts immigrant minorities in inferior social positions is inconsistent with a genuinely two-way process of integration. The demand that socially excluded immigrant minorities integrate thus becomes a demand that they assimilate, just one that is couched in the language of integration. Few explicitly defend assimilation, because if we want to achieve an integrated society, then it is fair to require that both minorities and majorities mutually adjust to achieve that social goal and unfair to require adjustment only of minorities. This interpretation treats social exclusion itself as incompatible with the process of integration and suggests that we should view the duty to integrate as a *conditional* duty that depends on dominant majority groups being credibly committed to doing their part in the process of integration.

This latter interpretation of the claim that the duty to integrate is unreasonably burdensome also supports a further step in the negative argument: even if it is a genuine duty, integration cannot be enforced as a social expectation because majorities lack the standing to blame socially excluded immigrants for failing to integrate. This further claim does not establish that socially excluded immigrants have no moral duty to integrate. Rather, it establishes that *even if* socially excluded immigrants have a moral duty to integrate, that duty cannot be enforced as a social expectation.

The basic idea here is that the claim that immigrant minorities ought to integrate is *second-personal* in nature. In Stephen Darwall's terms, second-personal claims come "with an RSVP attached": they make a demand of the addressee to act in a particular way or to account for their behavior to the speaker if they fail to do so.[53] The addressee of a justified second-personal claim is liable to be blamed if they fail to comply. Blaming is a paradigmatically *communicative* act that aims to make the addressee see the force of the shared moral reasons that the speaker presupposes in making a claim against them.[54] A social expectation is a generalized form of a second-personal claim in which those who uphold a social expectation treat those who fail to fulfill it as being liable to blame.

If integration is to be enforced as a social expectation, then not only must immigrant minorities be blameworthy for failing to integrate, but those who uphold the social expectation must also have standing to blame them for their failures. I have already suggested that socially excluded immigrant minorities do not have a duty to integrate and so are not liable to blame for failing to do so. But beyond this, my suggestion is that *even if* socially excluded immigrants do have a duty to integrate, such a duty cannot be enforced as a social expectation because members of dominant majority groups do not have standing to blame socially excluded immigrants who do not integrate. Regardless of whether or not socially excluded immigrants are blameworthy for not integrating, members of dominant majority groups who uphold the social expectation of integration are not, in Marilyn Friedman's terms, "blamer-worthy."[55]

In the literature on blame, two conditions for standing to blame have been identified: the *nonhypocrisy condition* and the *noninvolvement condition*.[56] The first suggests that those who have committed the same or a similar wrong to the target lack standing to blame. The second suggests that those who are in some way involved in the target's wrongdoing lack standing to blame. There is some disagreement about these conditions. For example, one disagreement concerns whether the nonhypocrisy condition is better explained by a lack of commitment to the relevant moral norm or by the idea that hypocrites reject the equality of persons by making an exception of themselves.[57] But these disagreements need not trouble us, because the social expectation that socially excluded immigrant minorities participate in social integration can be rejected on either

53 Darwall, *The Second Person Standpoint*, 40–41.

54 Fricker, "What's the Point of Blame?"

55 Friedman, "How to Blame People Responsibly," 272.

56 Todd, "A Unified Account of the Moral Standing to Blame."

57 For the former view, see Rossi, "The Commitment Account of Hypocrisy." For the latter view, see Fritz and Miller, "Hypocrisy and the Standing to Blame."

the grounds of the nonhypocrisy condition or the noninvolvement condition, whether we adopt the commitment account or the equality account of hypocrisy.

The nonhypocrisy condition says that those who have committed the same or a similar wrong to the target lack standing to blame. In contexts of social exclusion, dominant majority groups collectively fall foul of this condition, which means that social integration cannot be enforced as a social expectation. In a society in which immigrant minorities face social exclusion, dominant majority groups collectively uphold norms that put immigrant minorities at the bottom end of a pervasive hierarchy of esteem. This is inconsistent with the genuine participation of the dominant group in the process of integration. Integration involves reciprocal duties on the part of both immigrant minority and dominant majority groups. When dominant majority groups collectively uphold norms of social exclusion, they do not do their part in the process of integration. It is hypocritical of them to hold socially excluded immigrant minorities to the duty of integration when they themselves fail to fulfill the same duty. On the commitment account of hypocrisy, their social exclusion of immigrant minorities betrays their lack of commitment to the moral norm of integration. On the equality account of hypocrisy, those who hold immigrant minorities but not themselves to the moral norm of integration make an exception of themselves and so violate the moral equality of persons. Whichever account of hypocrisy we adopt, we can say that dominant majority groups collectively lack the standing to blame socially excluded immigrant minorities for failing to integrate. Since the social expectation of integration requires that dominant majority groups have standing to blame for failures to integrate, this means that the social expectation of integration cannot be enforced vis-à-vis socially excluded immigrants.

The noninvolvement condition says that those who are involved in the target's wrongdoing lack standing to blame. The notion of involvement is somewhat vague, but in the case at hand it can be rendered in the following way: dominant majorities are involved in the failure of socially excluded immigrant minorities to integrate because they have created the conditions in which discharging the duty to integrate is highly burdensome. Social integration is burdensome for socially excluded immigrants, who can expect to be exposed to stigma and hostility in their interactions with members of dominant majority groups. Collectively, dominant majorities are responsible for making it burdensome for socially excluded immigrants to discharge their duty of social integration. When socially excluded immigrants fail to discharge that duty as a result of those burdens, dominant majorities are involved in the failure to discharge the duty of social integration. And when dominant majority groups are involved in the failure of immigrant minorities to integrate by upholding

norms of social exclusion, those dominant majorities lack standing to blame immigrant minorities for failing to integrate.

The unreasonable burdens argument suggests that socially excluded immigrant minorities have a moral permission to form enclaves because they do not have a duty to integrate, such that they are not blameworthy for failures of integration. The standing to blame argument suggests that even if socially excluded immigrants do have such a duty, it cannot be permissibly enforced as a social expectation.

This standing argument leaves open the possibility that those who are not implicated in the social exclusion of immigrant minorities—other members of the socially excluded group, for example—might have standing to blame those who fail to integrate. It also leaves open the possibility that majorities might either acquire standing to blame by changing social conditions such that immigrants no longer face social exclusion or have standing to blame immigrants who are not socially excluded. As I have already suggested, I do not think that socially excluded immigrants do have a genuine duty to integrate. But if they ultimately do have such a duty, then it seems plausible to suggest that it would be other socially excluded immigrants (rather than dominant majorities who are implicated in social exclusion) who have standing to enforce that duty through social sanctions such as blame. If anyone has standing to blame, then it is others who are similarly situated vis-à-vis the problem of social exclusion. I take this to be a welcome implication of the argument from standing to blame.

## 3. OBJECTIONS

In this section, I consider two objections to my argument. The first objection says that because many immigrants—unlike members of other social groups—have chosen to enter a country voluntarily, they have waived their moral permission to form enclaves. The second objection says that because social integration has an important causal role in reducing prejudice, enclave formation may hinder the pursuit of relational equality.

### *3.1. Voluntary Immigration and Enclaves*

The first objection says that because immigrants have chosen to enter a country voluntarily, they have thereby waived their moral permission to form enclaves. The basic idea is that since those who have immigrated voluntarily have made a free choice to do so, they cannot reasonably expect to escape a duty to integrate within their host society. In his discussion of immigrant integration, Will Kymlicka makes a similar argument about cultural minority rights. On his view, immigrants "voluntarily relinquish" or "waive" their claims to "live and work in

their own culture."[58] If this is right, then voluntary immigrants may waive their moral permission to form enclaves. As Kymlicka recognizes, one limit to this argument is that it only applies to those who have actually made a voluntary choice to migrate. This means that it does not apply to either the children of first-generation immigrants or refugees.[59] This limits the scope of the objection. But as sociologists who study migration point out, enclaves are typically most pronounced among first-generation immigrants in any case.[60] So the objection may nonetheless still apply to a considerable range of cases.

Although Kymlicka does view integration as a two-way process and suggest that states should work to reduce prejudice and discrimination against immigrants, he does not suggest that social exclusion affects the duty to integrate.[61] The hypothetical example that he uses to motivate his argument that immigrants waive their claim to cultural minority rights—the emigration of a group of Americans to Sweden—involves no pervasive hierarchy of esteem with the immigrant group at the bottom. But in reality, many immigrants—even voluntary ones—face social exclusion in their new societies. Might this mean that they retain their moral permission to form enclaves? Kymlicka's argument may apply to those who do not face social exclusion—I take no stand on that question here—but when it comes to the socially excluded, the picture is quite different.

We can view the decision to immigrate as the decision to accept a kind of implicit contract. On this picture, immigrants accept the terms that the state offers to them when they decide to settle within a society. The duty to integrate is one contractual term to which immigrants sign up when they decide to migrate. Those who are forced to migrate cannot be said to have accepted the terms that the state offers—they have accepted the migration contract only under duress—but this does not apply to voluntary immigrants.

One reason we might think that even voluntary immigrants do not waive their moral permission to form enclaves by migrating is because they have an right to migrate. If would-be immigrants have a right to migrate, then they cannot be reasonably required to forgo their moral permission to form enclaves in order to exercise that right. In a related discussion of whether immigrants can consent to permanent alienage (i.e., denizenship without access to citizenship), Kieran Oberman argues that permanent alienage is wrongful not because would-be immigrants cannot consent to it but because they have a right to

58 Kymlicka, *Multicultural Citizenship*, 96.

59 Kymlicka, *Multicultural Citizenship*, 98–100, 215–16n19.

60 Portes and Manning, "The Immigrant Enclave."

61 Kymlicka, *Multicultural Citizenship*, 96.

migrate.[62] This means that their exercise of their right to migrate cannot be taken as evidence that they have accepted the terms of the migration contract that states have imposed upon them. As he puts it, "if a voluntary migrant has a right to immigrate, then one cannot infer a migrant's consent to the terms of her admission from the fact that she has chosen to migrate."[63] Similarly, we might think that voluntary immigrants do not waive their moral permission to form enclaves by migrating because they have a right to migrate independently of whether or not they waive that permission.

The main problem with this argument is that it requires us to accept a controversial premise: that would-be immigrants have a right to migrate. To say that this premise is controversial is not to say that it is mistaken, and I remain agnostic here on whether or not there is a right to migrate. But my defense of enclaves will have much broader reach if it does not require us to accept this controversial premise and is instead consistent with what Carens calls the "conventional view" of the political morality of immigration, according to which each state has a discretionary right to exclude would-be immigrants.[64]

We can reject the claim that socially excluded immigrants waive their moral permission to form enclaves when they migrate voluntarily even within the constraints of the conventional view. This is because the receiving state having a discretionary right to exclude would-be immigrants is consistent with there being moral constraints on the exercise of that right. Just as an employer who has no duty to hire anyone faces constraints on the kinds of criteria they can use to make hiring decisions and the kinds of terms they can put in their employment contracts, so too are there moral constraints on the state's exercise of its discretionary right to exclude would-be immigrants.[65] One such constraint is that states may not impose unfair terms within the migration contract. When they do so, such terms are morally unenforceable.

The requirement that socially excluded immigrants waive their moral permission to form enclaves as a condition of entry should be viewed as an unfair and thus morally unenforceable term in the migration contract. Michael Blake has recently argued that states may implement only those immigration policies that would-be immigrants can accept "without accepting their own moral inferiority."[66] On Blake's view, this rules out immigration policies that select according to race or religion. But it also rules out a migration contract—even

62 Oberman, "Immigration, Citizenship, and Consent."

63 Oberman, "Immigration, Citizenship, and Consent," 105.

64 Carens, *The Ethics of Immigration*, 10.

65 Carens, *The Ethics of Immigration*, 174–75.

66 Blake, *Justice, Migration, and Mercy*, 121.

an implicit one—that requires immigrants to accept their own social exclusion as a condition of entry. Such a contract is unfair because requiring those who face social exclusion to waive their moral permission to form enclaves is akin to requiring them to acquiesce to their own subordination. It says to would-be immigrants that they can enter only on the condition that they accept that their place is at the bottom of the social hierarchy of esteem and give up the right to use defense mechanisms to protect their self-respect. Some would-be immigrants might well prefer to accept the offer to migrate under such conditions rather than to forgo the option of migrating at all. But this is not a choice that it is fair to ask them to make. Even if would-be immigrants were to voluntarily accept such a contract, its unfairness means that it morally unenforceable. If my landlord puts in my rental contract that I am not allowed to jump on my own bed and refuses to negotiate on this term, then the appropriate response is to smile, sign the paperwork, and jump on the bed anyway. My landlord has no right to make such a demand of me, and no reasonable tenancy law would permit him to enforce his claim that I not jump on my own bed.[67] Neither do receiving societies have the right to require that immigrant minorities accept their position as moral inferiors. This explains why socially excluded immigrant minorities retain their moral permission to form enclaves, even if they have migrated voluntarily and even if states have a discretionary right to exclude.

### *3.2. Integration and Prejudice-Reduction*

A second objection to my argument is that enclaves for the excluded may close off promising avenues for achieving relational equality. The basic idea here is that, at least according to some important findings in social psychology, integration can play an important role in reducing prejudice. Integration thus has the potential to ameliorate the condition of social exclusion faced by groups such as immigrant minorities. But if enclaves for the excluded are permitted, then this avenue for achieving relational equality is foreclosed, or at least hindered.

In social psychology, the "contact hypothesis" suggests that patterns of interaction across group lines can reduce prejudice.[68] The basic idea is that positive interactions between members of different social groups can break down prejudice by broadening the boundaries of the perceived in-group, reducing reliance on stereotypes and defusing anxiety and antipathy about interacting with those from other social groups. An influential meta-analysis has found that the vast majority of empirical tests support the claim that positive intergroup

67 This analogy is inspired by a similar one used in Jubb, "Consent and Deception," 227.

68 The *locus classicus* is Allport, *The Nature of Prejudice*.

contact typically reduces prejudice.[69] In relation to immigration in particular, research has shown that positive contact can reduce anti-immigrant prejudice, particularly by reducing the perceived threat felt by nonimmigrants.[70] Positive contact has also been shown to reduce the influence of inegalitarian social norms and to reduce support for anti-immigrant and far-right parties.[71] In the context of US racial politics, Elizabeth Anderson draws on the contact hypothesis in her defense of integration, arguing that it plays a critical role in prejudice reduction.[72] Likewise, we might argue that the beneficial effects of integration for prejudice reduction mean that we should reject the claim that socially excluded immigrants have a moral permission to form enclaves, since forming enclaves hinders prejudice-reducing forms of intergroup contact.

The empirical premise in this argument does require some qualification, but it remains strong overall. In Gordon Allport's original articulation of the contact hypothesis, he argued that intergroup contact reduces prejudice only when four conditions are met: contact must be *frequent*, *cooperative*, *institutionally scaffolded*, and of *equal status*.[73] The weight of the empirical evidence now suggests that these are best viewed as mediating conditions that can magnify or diminish the prejudice-reducing effect of positive intergroup contact, not as necessary conditions for prejudice reduction.[74] The social environment in which contact takes place does make a difference to the effectiveness of intergroup contact, but the relationship between intergroup contact and prejudice reduction is fairly robust, even outside of experimental settings.[75] There are also some limits to the contact hypothesis: incidences of negative contact may increase prejudice, informal practices of resegregation can limit opportunities for contact outside of experimental conditions, and intergroup contact may also have a "sedative effect" on collective resistance by disadvantaged social

69 Pettigrew and Tropp, "A Meta-Analytic Test of Intergroup Contact Theory."

70 Meleady, Seger, and Vermue, "Examining the Role of Positive and Negative Intergroup Contact and Anti-Immigrant Prejudice in Brexit"; Schneider, "Anti-Immigrant Attitudes in Europe"; Savelkoul et al., "Anti-Muslim Attitudes in the Netherlands"; and McLaren, "Anti-Immigrant Prejudice in Europe."

71 Visintin et al., "Intergroup Contact Moderates the Influence of Social Norms on Prejudice"; Andersson and Dehdari, "Workplace Contact and Support for Anti-Immigration Parties"; and Savelkoul, Laméris, and Tolsma, "Neighbourhood Ethnic Composition and Voting for the Radical Right in the Netherlands."

72 Anderson, *The Imperative of Integration*, 123–27.

73 Allport, *The Nature of Prejudice*.

74 Pettigrew, "Intergroup Contact Theory"; and Pettigrew and Tropp, "A Meta-Analytic Test of Intergroup Contact Theory."

75 Lemmer and Wagner, "Can We Really Reduce Ethnic Prejudice Outside the Lab?"

groups.[76] This latter effect is particularly important, as it suggests that there may be a trade-off between collective resistance—which, as we have seen, can be important in preserving self-respect—and intergroup contact. Still, despite these limits, the weight of the empirical evidence does support the broad claim that positive intergroup contact tends to reduce prejudice.

But even if the empirical premise in this argument is sound, it does not show that socially excluded immigrants have a moral duty to participate in integration. There are two reasons to think that even if the empirical premise of this argument is sound, it does not put socially excluded immigrants under a duty to participate in integration.

First is simply that the process of integration remains burdensome for socially excluded immigrants. As I have already argued, the social exclusion that some immigrant minorities face makes the demand that immigrant minorities integrate particularly burdensome for them. Socially excluded immigrants who engage in social integration can expect to be exposed to stigma and hostility in their interactions with nonimmigrants. Integration also requires them to forego the protective benefits that they can get from enclaves in terms of maintaining their self-respect, and this point is only strengthened by the finding that social integration can also have a sedative on collective resistance. If my previous arguments to this effect are correct, then requiring social integration would still seem to impose an unreasonable burden on socially excluded immigrants.[77]

Second is that when social integration is viewed as a tool for prejudice reduction, then this means that its ultimate beneficiaries are socially excluded immigrants themselves. So far, I have treated the putative duty to integrate as a duty that is owed to members of the receiving society. But if we care about integration because its prejudice-reducing effects mean that it promotes relational equality, then the putative duty to integrate is ultimately a duty that is owed to those who are the victims of relational inequality: in this case, socially excluded immigrants themselves. This makes an important difference to the argument for integration; it means that the benefits of social integration are not something that members of the receiving society can demand of socially excluded immigrants. Instead, this conception of the putative duty to integrate puts socially excluded immigrants themselves in the position of being able to decide whether or not to release themselves from this duty. In other words, it is the case both that socially excluded immigrants have good reasons to object

76 McKeown and Dixon, "The 'Contact Hypothesis'"; Cakal et al., "An Investigation of the Social Identity Model of Collective Action and the 'Sedative' Effect of Intergroup Contact among Black and White Students in South Africa"; and Dixon, Durrheim, and Tredoux, "Beyond the Optimal Contact Strategy."

77 See Shelby, *Dark Ghettos*, 73–76.

to being required to participate in social integration and that their participation is not something that is ultimately owed to members of the receiving society.

Where does this leave us with respect to social integration? It may well be that without integration, a society of equals will remain only an ideal that cannot be fully realized. But at the same time, it may be unreasonably burdensome to require that socially excluded immigrant minorities participate in integration. Social exclusion both makes it the case that immigrant minorities have only limited moral duties to participate in integration and at the same time makes integration all the more important. This is ultimately why my defense of enclaves for the excluded is a pessimistic one. On this view, it is better from the point of view of relational equality if socially excluded immigrants participate in integration, but their participation in integration is supererogatory. This suggests that instead of treating integration as an expectation, we should treat it as an aspiration. The fact that many socially excluded immigrants do participate in integration, despite their lack of a duty to do so, should be a cause for celebration. But it is not something that can be reasonably required of them.

## 4. CONCLUSION

Immigrants are typically expected to participate in social integration in their receiving societies. But some immigrant minorities are subject to this expectation while at the same time being placed in an inferior social position in a pervasive hierarchy of esteem. In this paper, I have argued that those in this position—socially excluded immigrant minorities—have a moral permission to form enclaves, which means that they have only limited duties to participate in social integration. Positively, enclaves can have a protective function against the threats to self-respect involved in social exclusion. Negatively, social exclusion makes the putative duty to integrate unreasonably burdensome. And further, social integration cannot be justified as a social expectation because members of dominant majority groups lack the standing to blame socially excluded immigrant minorities for failures to integrate.

However, it is true that social integration is an important tool for combatting relational inequality. This makes my argument a pessimistic one: social exclusion both makes it the case that socially excluded immigrant minorities have only limited duties to participate in integration and makes it all the more important that they do so, if we are to achieve relational equality. We may hope that socially excluded immigrants integrate, and the fact that many do so may be a cause for celebration. But the integration of socially excluded immigrant minorities is not something that we can legitimately expect, and when socially

excluded immigrants do participate in integration, they are doing something supererogatory.

One attractive feature of this defense of enclaves is that it is *asymmetric*: it applies only to members of socially excluded groups and not to members of social groups who do not face social exclusion. These features of my account enable it to avoid yielding implausible judgments about other cases of enclave formation that do not meet these conditions. Consider, for example, affluent white Americans who cluster together in gated communities. Geographers and sociologists have pointed out that despite being facially neutral, gated communities enable affluent white Americans to engage in social closure by excluding minority groups.[78] This kind of enclave formation cannot be justified by my defense of enclaves. Because affluent white Americans do not face social exclusion, they do not have a justification for engaging in enclave formation on the basis of self-respect. My argument thus avoids the implausible conclusion that members of dominant majority groups have a moral permission to form enclaves.

One upshot of my argument is that debates about immigrant integration should be much more focused on the duties of members of receiving societies than on the duties of immigrants. It suggests that the onus is on members of dominant social groups who uphold hierarchies of esteem that put some immigrants in an inferior social position to change their behaviors. It is only when immigrant minorities do not face social exclusion that they can be held to the expectation that they should participate in social integration.[79]

*Utrecht University*
*j.r.g.draper@uu.nl*

## REFERENCES

Allport, Gordon W. *The Nature of Prejudice*. Addison-Wesley, 1954.

Anderson, Elizabeth. "Equality." In *The Oxford Handbook of Political Philosophy*, edited by David Estlund. Oxford University Press, 2012.

78 Le Goix and Webster, "Gated Communities."

79 I owe special thanks to Uğur Aytac, Jacob Barrett, Jelena Belic, Rebecca Buxton, Josette Daemen, Maxime Lepoutre, Andrew Mason, Alex McLaughlin, Kieran Oberman, Johan Olsthoorn, and Emily McTernan for helpful suggestions on this paper. I would also like to thank audiences at the Trust and the Integration of Immigrants workshop at the University of Oxford, the Good Integration workshop at the University of Tromsø, the UCL Colloquium in Legal and Political Theory, and the OZSW Workshop in Political Philosophy at Utrecht University for helpful feedback.

———. *The Imperative of Integration*. Princeton University Press, 2010.
Andersson, Henrik, and Sirus H. Dehdari. "Workplace Contact and Support for Anti-Immigration Parties." *American Political Science Review* 115, no. 4 (2021): 1159–74.
Bell, David Andreas, Marko Valenta, and Zan Strabac. "A Comparative Analysis of Changes in Anti-Immigrant and Anti-Muslim Attitudes in Europe: 1990–2017." *Comparative Migration Studies* 9, no. 1 (2021): 57.
Bicchieri, Christina. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, 2006.
Bird, Colin. "Self-Respect and the Respect of Others." *European Journal of Philosophy* 18, no. 1 (2010): 17–40.
Blake, Michael. *Justice, Migration, and Mercy*. Oxford University Press, 2020.
Bohman, James. *Public Deliberation: Pluralism, Complexity, and Democracy*. Massachusetts Institute of Technology Press, 1996.
Boxill, Bernard R. "Self-Respect and Protest." *Philosophy and Public Affairs* 6, no. 1 (1976): 58–69.
Bratu, Christine. "Self-Respect and the Disrespect of Others." *Ergo* 6, no. 13 (2019): 357–73.
Brownlee, Kimberley. "Freedom of Association: It's Not What You Think." *Oxford Journal of Legal Studies* 35, no. 2 (2015): 267–82.
Brubaker, Rogers. "Between Nationalism and Civilizationism: The European Populist Moment in Comparative Perspective." *Ethnic and Racial Studies* 40, no. 8 (2017): 1191–226.
Cakal, Huseyin, Miles Hewstone, Gerhard Schwär, and Anthony Heath. "An Investigation of the Social Identity Model of Collective Action and the 'Sedative' Effect of Intergroup Contact among Black and White Students in South Africa." *British Journal of Social Psychology* 50, no. 4 (2011): 606–27.
Carens, Joseph. *The Ethics of Immigration*. Oxford University Press, 2013.
———. "The Integration of Immigrants." *Journal of Moral Philosophy* 2, no. 1 (2005): 29–46.
Castells, Manuel. *The City and the Grassroots: A Cross-Cultural Theory of Urban Social Movements*. University of California Press, 1983.
Darby, Derrick, and Eduardo J. Martinez. "Making Identities Safe for Democracy." *Journal of Political Philosophy* 30, no. 3 (2022): 273–97.
Darwall, Stephen L. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Harvard University Press, 2006.
———. "Two Kinds of Respect." *Ethics* 88, no. 1 (1977): 36–49.
De Haas, Hein. *How Migration Really Works: A Factful Guide to the Most Divisive Issue in Politics*. Viking Press, 2023.
De Haas, Hein, Stephen Castles, and Mark J. Miller. *The Age of Migration:*

*International Population Movements in the Modern World*. Bloomsbury, 2019.
Dixon, John, Kevin Durrheim, and Colin Tredoux. "Beyond the Optimal Contact Strategy: A Reality Check for the Contact Hypothesis." *American Psychologist* 60 (2005): 697–711.
Draper, Jamie. "Gentrification and Everyday Democracy." *European Journal of Political Theory* 23, no. 3 (2024): 359–80.
Fraser, Nancy. "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy." *Social Text* 25–26 (1990): 56–80.
Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007.
———. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs* 50, no. 1 (2016): 165–83.
Friedman, Marilyn. "How to Blame People Responsibly." *Journal of Value Inquiry* 47, no. 3 (2013): 271–84.
Fritz, Kyle G., and Daniel Miller. "Hypocrisy and the Standing to Blame." *Pacific Philosophical Quarterly* 99, no. 1 (2018): 118–39.
Gorodzeisky, Anastasia, and Moshe Semyonov. "Not Only Competitive Threat but Also Racial Prejudice: Sources of Anti-Immigrant Attitudes in European Societies." *International Journal of Public Opinion Research* 28, no. 3 (2016): 331–54.
Haslanger, Sally. *Resisting Reality: Social Construction and Social Critique*. Oxford University Press, 2012.
Hay, Carol. "The Obligation to Resist Oppression." *Journal of Social Philosophy* 42, no. 1 (2011): 21–45.
Holtug, Nils. *The Politics of Social Cohesion: Immigration, Community, and Justice*. Oxford University Press, 2021.
Honneth, Axel. *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. Massachusetts Institute of Technology Press, 1995.
Hughes, David. "Braverman: Immigrants Living 'Parallel Lives' in Many UK Towns and Cities." *Independent*, October 3, 2023.
Joppke, Christian. "Beyond National Models: Civic Integration Policies for Immigrants in Western Europe." *West European Politics* 30, no. 1 (2007): 1–22.
Jubb, Robert. "Consent and Deception." *Journal of Ethics and Social Philosophy* 12, no. 2 (2017): 223–29.
Klarenbeek, Lea M. "Reconceptualising 'Integration as a Two-Way Process.'" *Migration Studies* 9, no. 3 (2021): 902–21.
Kolodny, Niko. *The Pecking Order: Social Hierarchy as a Philosophical Problem*. Harvard University Press, 2023.
Krishnamurthy, Meena. "Completing Rawls's Arguments for Equal Political

Liberty and Its Fair Value: The Argument from Self-Respect." *Canadian Journal of Philosophy* 43, no. 2 (2013): 179–205.
Kymlicka, Will. *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Clarendon Press, 1996.
Laborde, Cécile. *Critical Republicanism: The Hijab Controversy and Political Philosophy*. Oxford University Press, 2008.
Lægaard, Sune. "Unequal Recognition, Misrecognition and Injustice: The Case of Religious Minorities in Denmark." *Ethnicities* 12, no. 2 (2012): 197–214.
Le Goix, Renaud, and Chris J. Webster. "Gated Communities." *Geography Compass* 2, no. 4 (2008): 1189–214.
Lemmer, Gunnar, and Ulrich Wagner. "Can We Really Reduce Ethnic Prejudice Outside the Lab? A Meta-Analysis of Direct and Indirect Contact Interventions." *European Journal of Social Psychology* 45, no. 2 (2015): 152–68.
Levy, Jacob T. "Multicultural Manners." In *The Plural States of Recognition*, edited by Michel Seymour. Palgrave Macmillan, 2010.
Mansbridge, Jane. "Everyday Talk in the Deliberative System." In *Deliberative Politics: Essays on Democracy and Disagreement*, edited by Stephen Macedo. Oxford University Press, 1999.
Marcuse, Peter. "The Enclave, the Citadel, and the Ghetto: What Has Changed in the Post-Fordist US City." *Urban Affairs Review* 33, no. 2 (1997): 228–64.
Margalit, Avishai. *The Decent Society*. Harvard University Press, 1996.
Mason, Andrew. "The Critique of Multiculturalism in Britain: Integration, Separation and Shared Identification." *Critical Review of International Social and Political Philosophy* 21, no. 1 (2018): 22–45.
———. *Living Together as Equals: The Demands of Citizenship*. Oxford University Press, 2012.
McKeown, Shelley, and John Dixon. "The 'Contact Hypothesis': Critical Reflections and Future Directions." *Social and Personality Psychology Compass* 11, no. 1 (2017): e12295.
McLaren, Lauren M. "Anti-Immigrant Prejudice in Europe: Contact, Threat Perception, and Preferences for the Exclusion of Migrants." *Social Forces* 81, no. 3 (2003): 909–36.
McTernan, Emily. "The Inegalitarian Ethos: Incentives, Respect, and Self-Respect." *Politics, Philosophy and Economics* 12, no. 1 (2013): 93–111.
———. "Microaggressions, Equality, and Social Practices." *Journal of Political Philosophy* 26, no. 3 (2018): 261–81.
Medina, José. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. Oxford University Press, 2013.
Meleady, Rose, Charles R. Seger, and Marieke Vermue. "Examining the Role of Positive and Negative Intergroup Contact and Anti-Immigrant Prejudice in

Brexit." *British Journal of Social Psychology* 56, no. 4 (2017): 799–808.
Merry, Michael S. *Equality, Citizenship, and Segregation: A Defense of Separation*. Palgrave Macmillan, 2013.
———. "Equality, Self-Respect, and Voluntary Separation." *Critical Review of International Social and Political Philosophy* 15, no. 1 (2012): 79–100.
Miller, David. *Strangers in Our Midst: The Political Philosophy of Immigration*. Harvard University Press, 2016.
Modood, Tariq. *Multiculturalism: A Civic Idea*. Polity, 2007.
Nicholls, Walter. "Place, Networks, Space: Theorising the Geographies of Social Movements." *Transactions of the Institute of British Geographers* 34, no. 1 (2009): 78–93.
Oberman, Kieran. "Immigration, Citizenship, and Consent: What's Wrong with Permanent Alienage?" *Journal of Political Philosophy* 25, no. 1 (2017): 91–107.
Parekh, Bhikhu C. *Rethinking Multiculturalism: Cultural Diversity and Political Theory*. Harvard University Press, 2000.
Pettigrew, Thomas F. "Intergroup Contact Theory." *Annual Review of Psychology* 49, no. 1 (1998): 65–85.
Pettigrew, Thomas F., and Linda R. Tropp. "A Meta-Analytic Test of Intergroup Contact Theory." *Journal of Personality and Social Psychology* 90 (2006): 751–83.
Portes, Alejandro, and Robert D. Manning. "The Immigrant Enclave: Theory and Empirical Examples." In *Social Stratification, Class, Race, and Gender in Sociological Perspective*, edited by David B. Grusky. Routledge, 2001.
Quillian, Lincoln. "Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-Immigrant and Racial Prejudice in Europe." *American Sociological Review* 60, no. 4 (1995): 586–611.
Rawls, John. *A Theory of Justice*. Rev. ed. Harvard University Press, 1999.
Rossi, Benjamin. "The Commitment Account of Hypocrisy." *Ethical Theory and Moral Practice* 21, no. 3 (2018): 553–67.
Saeed, Amir. "Media, Racism, and Islamophobia: The Representation of Islam and Muslims in the Media." *Sociology Compass* 1, no. 2 (2007): 443–62.
Sangiovanni, Andrea, and Juri Viehoff. "Solidarity in Social and Political Philosophy." In *Stanford Encyclopedia of Philosophy* (Summer 2023). https://plato.stanford.edu/archives/sum2023/entries/solidarity/.
Savelkoul, Michael, Joran Laméris, and Jochem Tolsma. "Neighbourhood Ethnic Composition and Voting for the Radical Right in the Netherlands: The Role of Perceived Neighbourhood Threat and Interethnic Neighbourhood Contact." *European Sociological Review* 33, no. 2 (2017): 209–24.
Savelkoul, Michael, Peer Scheepers, Jochem Tolsma, and Louk Hagendoorn.

"Anti-Muslim Attitudes in the Netherlands: Tests of Contradictory Hypotheses Derived from Ethnic Competition Theory and Intergroup Contact Theory." *European Sociological Review* 27, no. 6 (2011): 741–58.

Schemmel, Christian. "Real Self-Respect and Its Social Bases." *Canadian Journal of Philosophy* 49, no. 5 (2019): 628–51.

Schneider, Silke L. "Anti-Immigrant Attitudes in Europe: Outgroup Size and Perceived Ethnic Threat." *European Sociological Review* 24, no. 1 (2008): 53–67.

Seglow, Jonathan. "Hate Speech, Dignity and Self-Respect." *Ethical Theory and Moral Practice* 19, no. 5 (2016): 1103–16.

Semyonov, Moshe, Rebeca Raijman, and Anastasia Gorodzeisky. "The Rise of Anti-Foreigner Sentiment in European Societies, 1988–2000." *American Sociological Review* 71, no. 3 (2006): 426–49.

Shelby, Tommie. *Dark Ghettos: Injustice, Dissent, and Reform*. Harvard University Press, 2016.

Soja, Edward W. *Seeking Spatial Justice*. University of Minnesota Press, 2010.

Todd, Patrick. "A Unified Account of the Moral Standing to Blame." *Noûs* 53, no. 2 (2019): 347–74.

Toole, Briana. "Recent Work in Standpoint Epistemology." *Analysis* 81, no. 2 (2021): 338–50.

Visintin, Emilio Paolo, Eva G. T. Green, Juan Manuel Falomir-Pichastor, and Jacques Berent. "Intergroup Contact Moderates the Influence of Social Norms on Prejudice." *Group Processes and Intergroup Relations* 23, no. 3 (2020): 418–40.

Wolff, Jonathan. "Fairness, Respect, and the Egalitarian Ethos." *Philosophy and Public Affairs* 27, no. 2 (1998): 97–122.

Young, Iris Marion. *Inclusion and Democracy*. Oxford University Press, 2002.

*Journal of Ethics and Social Philosophy*
VOL. 29, NO. 2 · JANUARY 2025
*https://doi.org/10.26556/jesp.v29i2.3481*


# VOTING, REPRESENTATION, AND INSTITUTIONS

## A CRITIQUE OF ELLIOTT'S DUTY TO VOTE

*Ben Saunders*

KEVIN J. ELLIOTT has recently offered a new institutional argument for a duty to vote, based on role obligations and the requirements of representation.[1] If certain groups do not vote, their interests may be neglected and/or misunderstood.[2] In contrast, voting promotes representative responsiveness. Thus, Elliott argues, universal turnout is *ordinarily necessary* for fair representation (913). Since he holds that citizens have a duty to do what is necessary for the proper functioning of representative institutions, in virtue of occupying the office of citizen, it follows that they have a duty to vote in elections.[3] However, this duty need not be absolute; presumably its stringency will depend on the significance of the election.

This argument is original and important, since it grounds the duty to vote on the internal logic of democratic institutions rather than on more basic moral duties such as samaritanism or fair play (902). Moreover, unlike some previous arguments, it purports to explain why citizens are under a duty to *vote* rather than participating in other ways (916–20). Voting is not simply one way among others to discharge some more general duty, such as contributing to the societal good. Rather, voting is special because it uniquely authorizes representatives (914). Therefore, voting is an "institutionally specific need of electoral representative democracy" (918). Without universal voting, Elliott argues, electoral representation will work less well.

The "necessity" that Elliott claims for universal turnout is not strict, logical necessity but rather practical or realistic necessity (919). He holds that our thinking about political institutions (and associated duties) should be guided

1 Elliott, "An Institutional Duty to Vote" (hereafter cited parenthetically).

2 Elliott recognizes that there are many different theories of representation (906–11). Both his arguments and mine are supposed to be neutral between these various accounts. For ease of exposition, I will speak throughout of interests as what should be represented.

3 I take this to include both national and local elections, though this could be rather demanding. See Rusavuk, "Which Elections?"

by what is likely or typical rather than by rare or exceptional occurrences. So, he dismisses certain theoretical possibilities, such as comprehensive altruism, as artificial and largely irrelevant. There is certainly some merit to this approach. Nonetheless, I find Elliott's case for the necessity of electoral turnout unconvincing. Indeed, he seems to undermine his own argument when he suggests that representatives will cater to uninformed voters and even those who spoil their ballots. This assumes that would-be representatives can identify the true interests of these voters and that they will be motivated to seek their votes.

Elliott identifies informational and motivational problems as two reasons why the interests of nonvoters may be neglected (918). First, if people do not vote, then their interests may not be properly understood, even by those acting in good faith. Second, if members of a certain group are unlikely to vote, then this reduces representatives' electoral incentives to respond to their known interests. These problems are self-reinforcing. If certain groups are less likely to vote, then representatives are less likely to cater to their interests, which is likely to further alienate them (914–15). However, while these are real problems, it is not clear that they are as serious as Elliott suggests—or that the duty to vote helps to overcome them.[4]

It has long been recognized that members of one group may not understand the perspectives or interests of other groups. Elliott's innovation is to argue that the mere *right* to vote is not enough; all social groups must actually vote. If they do not, then their interests may not be properly represented. Of course, Elliott does not claim that universal voting is sufficient for accurate representation. But understanding why it is not sufficient may lead us to question his claim that it is ordinarily necessary.

People vote as they do for a variety of reasons. We cannot assume that what people vote for is always in their interest—or even perceived by them as being so. At the very least, information about voting patterns needs to be supplemented, for instance with public opinion research, to identify people's wants. Elliott acknowledges a place for public opinion research (915), but only in connection to those who cast blank or spoiled ballots. Though he argues that even spoiled ballots convey something valuable about dissatisfaction, he does not explicitly say whether spoiling one's ballot satisfies the duty to vote. In either case, I find it hard to see how spoiled ballots help to overcome the information problem. They may express dissatisfaction—and, in this respect, a spoiled ballot may be clearer than simply staying at home, which might be dismissed as apathy or indifference—but they do not tell us *what* voters are

4 If voting will not secure fair representation, it is unclear why people should do it, even if it is necessary. See Saunders, "Against Detaching the Duty to Vote."

dissatisfied with or what it would take to satisfy them. Thus, it is unclear how those seeking election could win over these discontented voters.

The informational problems are exacerbated even further once we relax the assumption that voters are well informed.[5] Elliott argues that it matters little whether voters are informed, since electoral candidates cannot rely on voters to be uninformed (921). Their uncertainty, he says, gives them reason to act *as if* their constituents are well informed. I am not entirely clear why this should be so. First, although it is *possible* that uninformed voters may become more informed, this is often unlikely, since acquiring information is costly and voters lack incentive to do this.[6] Thus, politicians might reasonably expect ignorance to persist. Further, to the extent that representatives are motivated by electoral incentives, they will presumably do what they think will win votes. If voters are not perfectly informed, it is possible that they will vote for parties or policies that are not in—and perhaps even contrary to—their true interests. Unscrupulous politicians might take advantage of voter ignorance to serve their own ends.[7] Even if they are not seeking to exploit voter ignorance, they might instead act based on their best guesses about what people are likely to vote for rather than what they think is truly best for voters.

Let us grant Elliott's claim that representatives can be incentivized to promote the public good even when electoral incentives are uncertain. This presupposes that representatives can anticipate what well-informed voters would want, ahead of their voting, and even if those voters have previously voted in an ill-informed manner. But this implies that the information problems Elliott alludes to can be overcome after all. If this is so, then it significantly weakens the argument that universal voting is needed in order to provide representatives with information about citizens' interests.

Of course, Elliott does not say that one person can never accurately recognize another's interests. He may concede that representatives can to some extent identify the interests of citizens independently of their voting behavior yet maintain that this process will be more accurate and reliable when citizens vote than when they do not. However, we have already seen reasons to question

5 Elliott argues that citizens have a duty to vote, but this does not require them to vote well. This contradicts both those who argue for a positive duty to vote well (e.g., Klijnman, "An Epistemic Case for Positive Voting Dutiesm"; and Maskivker, "Merely Voting or Voting Well?") and those who argue for a negative duty not to vote badly (e.g., Brennan, "Polluting the Polls").

6 Klijnman, "An Epistemic Case for Positive Voting Duties," 77.

7 It should be noted that representatives do not merely represent pre-existing interests; as noted by Disch, they sometimes play a creative role in constructing constituencies and interests ("The 'Constructivist Turn' in Political Representation").

the helpfulness of voting here, especially if the votes in question include spoiled ballots and ill-informed votes that might actually be contrary to the citizens' true interests. If voting is an unreliable indicator of people's interests, then it is not clear how helpful universal turnout is in overcoming information problems and even less obvious that it is necessary to doing so.

This argument also threatens to undermine the motivational problem. Elliott suggests that representatives lack incentive to appeal to groups that do not vote (915–16). Yet when addressing the problem of citizen ignorance, he maintains that the "threat of electoral sanction works to a significant degree even when the sanction is uncertain" (921). If this is so, then the electoral sanction should still be effective when the uncertainty concerns turnout rather than informed voting. Even if a certain social group are known not to have voted in the past, representatives cannot count on their continued abstention. If there is a chance that some salient news story or unforeseeable event can overcome information deficits, then there is similarly a chance that something could mobilize previous nonvoters to vote. And if uncertainty leads to representative responsiveness in the one case, presumably this will also apply to the other. Thus, representatives might have incentive to act *as if* their constituents are likely to vote, whether or not this is actually the case.

Ideally perhaps, citizens should be attentive to politics and at least prepared to vote.[8] However, this is not necessary so long as political actors *believe* this to be the case. The mere *threat* of voting may be incentive enough to produce responsive representation. Hence, occasional nonparticipation need not undermine the functioning of the representative system so long as politicians cannot rely on this nonparticipation continuing. The problems of underrepresentation that Elliott points to arise only when nonparticipation goes beyond this, becoming habitual and expected (908).

Elliott might respond that nonparticipation can usually be predicted because political participation is habitual.[9] Thus, those who have not voted in the past are unlikely to vote in the future. However, these habits are not unbreakable. Since older people are generally more likely to vote than younger people, it must be that some nonvoters become voters as they age. Moreover, while some people may be habitual nonvoters, others may be occasional voters.[10] These

8 Tsoi defends a duty of attentiveness, without requiring people to vote ("You Ought to Know Better"). Elliott also emphasizes the importance of attentiveness, though he suggests compulsory voting as a means to promote this ("Aid for Our Purposes"). For criticism of this argument, see Pedersen et al., "Nudging Voters and Encouraging Pre-commitment."

9 I thank an anonymous referee for suggesting this response.

10 Bagozzi and Marchetti, "Distinguishing Occasional Abstention from Routine Indifference in Models of Vote Choice," 278; and Rapeli et al., "When Life Happens," 1244.

occasional voters may or may not vote depending on factors such as election campaigns or even the weather on election day.[11] Consequently, turnout is hard to predict. This uncertainty creates incentive to respond to potential voters. If existing representatives ignore these people, there is a danger that some political entrepreneurs will succeed in mobilizing them.[12]

Of course, representatives might still be *more* responsive to those they think more likely to vote. Thus, unequal participation can still lead to unequal representation. But still, this uncertainty undermines Elliott's claim that nonvoters *must* be ignored (916). At least in certain conditions, it might be easier to mobilize nonvoters to vote than to change how existing voters vote.[13] Therefore, politicians may have more need to be responsive to the interests of nonvoters (who may become voters) than to the interests of uninformed voters (who may become informed).

The incentives that representatives face are also influenced by institutional design.[14] Representatives are generally accountable to particular constituencies, so we can shape their incentives by (re)drawing these constituencies. Suppose members of a certain social group are less likely to vote than other groups and this threatens their substantive representation for the reasons Elliott suggests. Universal turnout is not the only solution. Another possibility is to give the group in question its own electoral constituencies. This is not simply another of those fanciful theoretical proposals that Elliott dismisses as unrealistic and irrelevant (912). Something like this has been done, for instance in New Zealand, which has separate constituencies for its indigenous Māori communities.[15] If constituencies are drawn in proportion to group size, then the group in question is guaranteed representation proportional to its numbers, even if turnout in these constituencies is lower than elsewhere. To be sure, such proposals face familiar difficulties. There are dangers of essentialism and legitimate worries that a majority within the group will dominate internal

11 Damsbo-Svendsen and Hansen, "When the Election Rains Out and How Bad Weather Excludes Marginal Voters from Turning Out"; Hillygus, "Campaign Effects and the Dynamics of Turnout Intention in Election 2000"; and Niven, "The Mobilization Solution?"

12 De Vries and Hobolt, *Political Entrepreneurs*, 219–20.

13 As an anonymous reviewer observes, an uninformed nonvoter will face two costs. It might be too costly for them to become an informed voter. But it does not follow that they should become an uninformed voter. See Maskivker, "Merely Voting or Voting Well?"; and Saunders "Against Detaching the Duty to Vote."

14 Given Elliott's realist objection to moralism (905), it is ironic that he focuses on individual duties rather than on system/institutional reform. For a critique of such approaches to participation, see Junn, "Diversity, Immigration, and the Politics of Civic Education."

15 McLeay, "Political Argument about Representation."

minorities. However, this example shows that representation of certain groups can be achieved, even with nonuniversal and indeed uneven turnout.

This does not necessarily undermine Elliott's arguments in other contexts, but it does at least show that the institutional duty to vote is contingent on institutional design. Elliott might respond that the institutional duty still applies to most familiar democratic systems, since arrangements like New Zealand's are unusual. However, constituency formation significantly affects group representation. Geographically concentrated groups are likely to be well represented, whereas dispersed groups are less likely to be adequately represented. In some cases, institutional design reduces the need to vote, while in others (e.g., safe seats) it reduces the effectiveness of voting. In both cases, this threatens to undermine the institutional argument for a duty to vote.

Further, universal turnout may sometimes be problematic, for instance if it exacerbates majority domination. If everyone votes, then the majority of votes will always reflect the majority group in society. This can mean that a relatively indifferent majority triumphs over a more affected minority. In contrast, if turnout is less than universal, the minority have some chance of getting their way, because they may be more likely to vote. Differential turnout between groups may track different stakes, in a manner approximating proportional influence.[16] Admittedly, this is unlikely to reflect stakes perfectly. In practice, there are other reasons (besides being less affected) explaining why some groups are less likely to vote. Nonetheless, universal turnout, at least when combined with equal votes and majority rule, is not necessarily the right way to strike an appropriate balance between different interests either.

I would concede that representative institutions might function better if citizens voted *well*—for instance, if they cast informed votes. However, Elliott defends a duty to vote rather than a duty to vote well (920–21). This includes casting ill-informed votes and possibly even spoiled ballots. It is not clear to me how this is conducive to the excellence of the representative system. These votes do not make it any easier to identify citizens' true interests (perhaps the reverse), nor do they give politicians incentives to promote the social good (again, possibly the reverse). Indeed, I am tempted by the stronger claim that representative democracy may function better *without* such votes.[17] Certainly, these votes are not necessary for its proper functioning. Representatives have other, possibly more reliable ways of identifying what people want and what is good for them. Moreover, they have incentives to respond, so long as there is a

16 Brighouse and Fleurbaey, "Democracy and Proportionality"; and Saunders, "The Democratic Turnout 'Problem,'" 317.

17 Brennan suggests a duty not to vote badly. See Brennan, "Polluting the Polls." For criticism of this argument, see Arvan, "People Do Not Have a Duty to Avoid Voting Badly."

credible threat of electoral sanction. This requires only that people *might* vote in future. Neither universal turnout nor a universal duty to vote is necessary.[18]

*University of Southampton*
*b.m.saunders@soton.ac.uk*

## REFERENCES

Arvan, Marcus. "People Do Not Have a Duty to Avoid Voting Badly: Reply to Brennan." *Journal of Ethics and Social Philosophy* 5, no. 1 (2011): 1–6.

Bagozzi, Benjamin E., and Kathleen Marchetti. "Distinguishing Occasional Abstention from Routine Indifference in Models of Vote Choice." *Political Science Research and Methods* 5, no. 2 (2017): 277–94.

Brennan, Jason. "Polluting the Polls: When Citizens Should Not Vote." *Australasian Journal of Philosophy* 87, no. 4 (2009): 535–50.

Brighouse, Harry, and Marc Fleurbaey. "Democracy and Proportionality." *Journal of Political Philosophy* 18, no. 2 (2010): 137–55.

Damsbo-Svendsen, Soren, and Kasper M. Hansen. "When the Election Rains Out and How Bad Weather Excludes Marginal Voters from Turning Out." *Electoral Studies* 81 (2023): 1–11.

De Vries, Catherine E., and Sara B. Hobolt. *Political Entrepreneurs: The Rise of Challenger Parties in Europe*. Princeton University Press, 2020.

Disch, Lisa. "The 'Constructivist Turn' in Political Representation." *Contemporary Political Theory* 11, no. 1 (2012): 114–18.

Elliott, Kevin J. "Aid for Our Purposes: Mandatory Voting as Precommitment and Nudge." *Journal of Politics* 79, no. 2 (2017): 656–69.

———. "An Institutional Duty to Vote: Applying Role Morality in Representative Democracy." *Political Theory* 51, no. 6 (2023): 897–924.

Hillygus, D. Sunshine. "Campaign Effects and the Dynamics of Turnout Intention in Election 2000." *Journal of Politics* 67, no. 1 (2005): 50–68.

Junn, Jane. "Diversity, Immigration, and the Politics of Civic Education." *Political Science and Politics* 37, no. 2 (2004): 253–55.

Klijnman, Carline. "An Epistemic Case for Positive Voting Duties." *Critical Review* 33, no. 1 (2021): 74–101.

Maskivker, Julia. "*Merely* Voting or Voting *Well*? Democracy and the

18 I first presented preliminary doubts about Elliott's argument in July 2023 at a workshop on democratic theory in Southampton, organized by William Chan. I thank that audience, especially David Owen, for comments and questions. I am also grateful to two anonymous reviewers for feedback on previous versions of the manuscript.

Requirements of Citizenship" *Inquiry* (forthcoming). Published online ahead of print, May 5, 2023. https://doi.org/10.1080/0020174X.2023.2208181.
McLeay, E. M. "Political Argument about Representation: The Case of the Maori Seats." *Political Studies* 28, no. 1 (1980): 43–62.
Niven, David. "The Mobilization Solution? Face-to-Face Contact and Voter Turnout in a Municipal Election." *Journal of Politics* 66, no. 3 (2004): 868–84.
Pedersen, Vikki M. L., Jens Damgaard Thaysen, and Andreas Albertsen. "Nudging Voters and Encouraging Pre-commitment: Beyond Mandatory Turnout." *Res Publica* 30, no. 2 (2024): 267–83.
Rapeli, Lauri, Achillefs Papageorgiou, and Mikko Mattila. "When Life Happens: The Impact of Life Events on Turnout." *Political Studies* 71, no. 4 (2023): 1243–60.
Rusavuk, Andre L. "Which Elections? A Dilemma for Proponents of the Duty to Vote." *Res Publica* 30, no. 3 (2024): 547–65.
Saunders, Ben. "Against Detaching the Duty to Vote." *Journal of Politics* 82, no. 2 (2020): 753–56.
———. "The Democratic Turnout 'Problem.'" *Political Studies* 60, no. 2 (2012): 306–20.
Tsoi, Siwing. "You Ought to Know Better: The Morality of Political Engagement." *Ethical Theory and Moral Practice* 21, no. 2 (2018): 329–39.

*Journal of Ethics and Social Philosophy*
VOL. 29, NO. 2 · JANUARY 2025
*https://doi.org/10.26556/jesp.v29i2.3556*


# COMMITTING TO PARENTHOOD

*Nicholas Hadsell*

How do adults acquire the right to parent a child? In *Parenting and the Goods of Childhood*, Luara Ferracioli proposes a moral commitment account of parenthood: "The parental role is best undertaken by those who *morally* commit to pursuing a parent-child relationship with a particular child."[1] In Ferracioli's defense of the moral commitment account, she claims it can accommodate worries about whether ambivalent gestating parents count as moral parents (they should) and whether it licenses parental proliferation (it should not). Here, I argue these worries are more worrisome than Ferracioli lets on.

## 1. WHAT IS THE MORAL COMMITMENT ACCOUNT?

### *1.1. Moral Commitment*

Let us start with what Ferracioli means by *moral commitment*. In her view, "moral commitments are commitments that persons make to morally valuable projects and relationships partly due to their recognition that such projects and relationships are of great value" (39). Ferracioli sees humanitarian work as a paradigm case of moral commitment, as those who engage in this work do so because they see charitable aid as a project that promotes value by raising well-being and respects value by serving other human beings with dignity. We can get more specific. In Ferracioli's view (40), *S* is morally committed to *Y* if and only if:

1. *S* is motivated by recognizing the value of *Y,*
2. *S* expresses her recognition of *Y*'s value through moral actions, and
3. *S* avoids expressions that violate stringent moral requirements.

The first condition says we must be sufficiently motivated by the right reasons for our commitment to a project or relationship to count as a moral commitment. If I am committed to helping those in poverty, my commitment is not moral if I am only motivated by how this would look on my resume and not

1 Ferracioli, *Parenting and the Goods of Childhood*, 30 (hereafter cited parenthetically).

by the well-being of those in poverty. I might be motivated to help for both reasons, but if the latter reason plays an insufficient role in my motivation, this will not count as a moral commitment. So, for childrearing to count as a moral commitment, the parent must be sufficiently motivated to rear her child by recognizing the value of the parent-child relationship (43).

Ferracioli's second condition requires us to *act* on the value recognition in the first condition for our commitment to count as moral. While we may generally act on our recognition of some valuable project or relationship, there are many cases in which we fail to do so. Ferracioli's example is instructive: "Paul might genuinely value the lives of poor people in the developing world and yet end up failing to donate to charity because he is too busy with his other projects" (41). Paul recognizes the value of charity but fails to act on that recognition due to his other tasks. So, our moral commitment to a valuable project or relationship must be a *commitment*: we must express ourselves by acting toward the project or relationship. Within the parent-child relationship, the parent acts on her recognition by adequately promoting and protecting her child's interests. If she only recognizes the value of the relationship but fails to act on it, she is not morally committed to raising her child.

The last condition is a basic constraint: moral commitments cannot happen at the expense of other moral requirements. Moral commitments should not lead us to do seriously immoral things. Even if charitable giving is worthwhile, we should not steal from our friends to pursue this project (42). This means that parents generally cannot promote and protect their children's interests by violating other stringent moral requirements. For example, adults who might otherwise meet their children's needs but conceive through sexual assault will not count as morally committed to the parent-child relationship (42).

So for *S* to morally commit to parenting a child, she must (1) be sufficiently motivated to take on the relationship by a recognition of its value, (2) act toward the child in ways that appropriately express her recognition of the relationship's value, and (3) avoid violating any stringent moral requirements (without good reason).[2]

### *1.2. Moral Commitments Beget Moral Rights*

On the moral commitment account, if we have an adult who satisfies the conditions of moral commitment to a child, then we also have an adult with the moral right to raise that child. The reason a moral right to parent follows from a parent's moral commitment to a child is that "a morally committed parent is necessarily a good parent . . . because a morally committed parent is necessarily

2 Ferracioli, *Parenting and the Goods of Childhood*, 43.

robustly disposed to take on the steps required for her child's life to go well" (43). That is, the moral right to parent follows the moral commitment to parent because morally committed parents will *care for their child reliably*, while parents without this disposition are not as reliable.[3]

Whereas other theories of parenthood (e.g., voluntaristic or causal accounts) do not guarantee that good parents will raise children, the virtue of Ferracioli's account is that it has such a guarantee. If a child is not raised by good parents, then they are not morally committed and therefore do not count as adults who have the moral right to raise their children (43–44). Parents can fail to be good by failing to be morally committed to their child in some way: they could fail to be sufficiently motivated (condition 1), fail to express their sufficient motivation in actions (condition 2), or fail to avoid violating their other moral requirements (condition 3). However, provided adults meet these conditions, they are moral parents.

### *1.3. Two Aspirations*

Ferracioli has two aspirations for the moral commitment account. First, the view should be *monistic*: it should "locate the grounds of moral parenthood in only one essential feature" (32). For any case of moral parenthood, the moral commitment account requires that the parent's moral commitment is the *only* thing that explains why that parent is the moral parent. In section 2, I will raise worries about whether the moral commitment account can deliver on this aspiration in cases where ambivalent procreators seem like moral parents while they also seem to lack a moral commitment. Second, the moral commitment account should explain an *exclusionary* moral right to raise a child: it "explains why other nonstate agents have a moral duty of noninterference and must respect the decisions undertaken by the moral parent" (30). In section 3, I will raise worries about whether the moral commitment account can explain

3 An anonymous reviewer asks why the parent's commitment entails a right. In Ferracioli's words, "Why should the moral right to parent attach to the morally committed parent? The reason is simple: a morally committed parent is necessarily a good parent." She follows up her discussion of this fact with: "It should therefore be clear that [the moral commitment account] will be quite well placed to comply with the aims of a dual-interest theory of the family," which is one of the desiderata Ferracioli is after in constructing an account of moral parenthood (43–44). So, as I see things, the reason the moral commitment entails a moral right to parent is that it secures the child's interests (because it is good for her to be raised by someone robustly disposed to care for her) and the parent's interests (because it is good for her to participate in what she sees as a deeply valuable relationship). Anyone not antecedently committed to the dual-interest theory will not find this persuasive, but because Ferracioli takes it on, I will too.

these exclusionary claims in cases where a third party wants to raise a child who already has a morally committed parent.[4]

## 2. AMBIVALENT PROCREATORS

The first worry I have about the moral commitment account is that it delivers counterintuitive results for ambivalent mothers and third parties. Consider the following case from Benjamin Lange.

> *Ambivalent Procreator*: Ann is in her second trimester and … has not decided yet whether she wants to rear the child with whom she is pregnant. By contrast, her friend Frank emotionally invests in and supports Ann's pregnancy and, without Ann's knowledge, forms the intention to parent the child himself.[5]

Who counts as morally committed to this child? Plausibly, Frank is morally committed, while Ann is not. Whereas Ann is ambivalent, Frank (1) is sufficiently motivated by a recognition of the value of a relationship with Ann's child, (2) acts on that commitment by helping Ann in her pregnancy, and (3) avoids violating any stringent moral requirements.[6] But if Frank is morally committed while Ann

4 Thanks to Anne Jeffrey for suggesting I make explicit how the worries below target the moral commitment account.

5 Lange, "A Project View of the Right to Parent," 15.

6 An anonymous reviewer says Ferracioli might deny Frank has a relationship with Ann's fetus. This is not to say third parties cannot have relationships with fetuses. Ferracioli herself notes that nongestating parties can count as morally committed "by supporting the gestating parent with the costs and hardships of pregnancy [or] preparation for taking up the parental role after birth" (*Parenting and the Goods of Childhood*, 45). Instead, the anonymous reviewer claims an asymmetry between third parties like Frank and intentional, nongestating co-procreators. Perhaps intentional co-procreators owe it to each other to facilitate engagement with the fetus in pregnancy because they embarked on this parental project together. This does not mean third parties have the standing to demand gestating parents make their bodies accessible, for this seems incompatible with the parents' bodily rights. But these third parties may nonetheless legitimately expect the gestating parent to give them opportunities to commit to the fetus they procreated together. Frank, however, is not Ann's co-procreator, which means she does not owe him any engagement with the fetus.

My response is twofold. First, even if we grant the asymmetry, this case is one in which Ann has allowed Frank's engagement anyway. If she has allowed him to engage with the fetus, it is beside the point whether she had the right to exclude him from doing so. Now that she has allowed Frank the space to engage, he satisfies the conditions of the moral commitment account and is thereby the moral parent. Second, the asymmetry relies on moral commitments having the power to exclude others from interfering with them; this is presumably why nongestating, intentional co-procreators can expect engagement with the fetus while third parties like Frank cannot. However, there are two problems with this. First, it assumes

is not, then we get a strange result: third parties can acquire a moral right to rear a child without the permission of ambivalent gestational parents. I imagine most would reject this strange result; however, if the moral commitment account is monistic, and Ann is the child's moral parent without a moral commitment, something else must explain why she is the moral parent and Frank is not.

However, Ferracioli denies that a third party like Frank should count as the moral parent and insists that ambivalent parents like Ann should count as morally committed:

> Gestating parents who are somewhat ambivalent about becoming parents can count as morally committed if, by the time the child is born, they have come to value the relationship with their newborn and want to maintain that relationship. After all, the mere act of gestation will be considered a form of recognition of the value of the future child, so long as the gestating parent does not actively harm the fetus by engaging in behavior that is clearly detrimental to its healthy development. (45)

The idea is this: ambivalent parents can still be moral parents by choosing not to harm their gestating child. This choice counts as a moral commitment because it expresses the parent's recognition of the value of her relationship with that child.

But there is a problem. Choosing not to harm one's gestating child does not necessarily count as a moral commitment in the same way that my choosing not to harm any third party does not necessarily count as a moral commitment. After all, we can abstain from harming others for various reasons that have nothing to do with recognizing the value of a relationship with them. My choosing to avoid harming a random bypasser on the street may result from my not giving them a second thought or being in a hurry to get somewhere else.[7]

When we consider Ann, things are no different. Recall that the moral commitment account says, "For a moral commitment to a particular parent-child relationship to arise, parents have to actually recognize the moral value of their unique paternalistic relationship with a particular child, and be sufficiently moved by *that* reason" (43). While many gestating parents act for this reason, Ann is not acting for this reason, which means she is not morally committed to the child. Of course, as Ferracioli points out, ambivalent parents like Ann may "choose not to have an abortion, decide not to take active steps to harm the fetus, and seek

---

Ann is the moral parent with the power to exclude, and I am arguing in this section that she is not the moral parent according to the moral commitment account. Second, even if Ann is the moral parent, I argue in section 3 below that the moral commitment account cannot explain how moral parents have the power to exclude others from inserting themselves into parent-child relationships. Either way, the asymmetry spells no trouble here.

7 Thanks to Anne Jeffrey for helping me sharpen this paragraph.

medical treatment and support during pregnancy" (44). While there are choices that secure the gestating child's biological interests, these are also choices Ann can make without any recognition of the value of a relationship with her child. When she does these things, perhaps she is simply doing what she thinks is right or conforming to what others expect her to do. Whatever Ann's reasons are, though, if one of them is not a recognition of the value of the parent-child relationship, she is not morally committed and is therefore not the moral parent.

So, how do we explain Ann's moral parenthood, something she seems to possess despite the preceding objection to the moral commitment account? If the moral commitment account cannot explain Ann's moral parenthood, then there must be some other normative feature—consent, causation, intentions, or something else—that explains our intuition that she is a moral parent. But if Ann's moral parenthood is explained by one of these other features and not a moral commitment, then the moral commitment account fails in its aspiration to be a monistic account—i.e., it does not hold that only one normative feature explains moral parenthood in every case.[8]

Now, one could appeal to further views in the metaphysics of pregnancy that complicate this verdict on Ann's moral parenthood. Fully evaluating this strategy's prospects is beyond this paper's scope; nonetheless, I want to give a brief cautionary note about this strategy through an example. Suppose a view says pregnancy is unique because there is no parent-child relationship yet for the parent to enjoy.[9] If true, perhaps the conditions for the expression of value the moral commitment account requires need some relaxation. Whereas the moral commitment of a parent whose child is already born will be extensive, a pregnant parent may still count as morally committed to her fetus so long as she meets her fetus's biological interests.[10]

Unfortunately, this view does not justify Ann's moral parenthood on the moral commitment account. After all, moral commitments to anything—relationships or projects—*require* a recognition of the value of the object of one's commitment, and the whole point of Ann's case is that she is not acting on or recognizing any value concerning her fetus. As long as the value of her fetus plays no role in her reasons for continuing the pregnancy, she is not morally

8 This outcome is not necessarily a reason to reject the moral commitment account simpliciter; plausible accounts of moral parenthood are pluralistic insofar as they allow multiple normative features to explain it. See, e.g., Bayne and Kolers, "Toward a Pluralistic Account of Parenthood." Ferracioli might just need to drop the monistic aspiration for the moral commitment account.

9 Thanks to an anonymous referee for raising this suggestion.

10 This is Ferrcaioli's own view, though she never claims it is an entailment of this particular view of pregnancy. Ferracioli, *Parenting and the Goods of Childhood*, 159.

committed. Pregnancy may be a case in which the conditions for moral commitment need relaxation, but the conditions cannot be so relaxed that they allow a moral commitment even when the agent does not satisfy two out of the three necessary conditions for a moral commitment—i.e., she has (1) no recognition of value, which means (2) she has no recognition of value to express in morally good actions. So, in short, the caution is this: no matter the metaphysical view of pregnancy on offer, moral commitments still need to be moral commitments.[11]

## 3. PARENTAL PROLIFERATION

The second worry I have about the moral commitment account is that it cannot explain why adults should not be able to insert themselves into existing parent-child relationships. Consider this next case.

> *The Prodigy*: Billy, a five-year-old prodigy, goes viral for playing "Bohemian Rhapsody" on the piano. Hundreds of adults across the country desire to parent Billy even though Billy already has two morally committed parents who do not want to co-parent with his fans.

11 An anonymous reviewer claims pregnancy does not create any new problem for Ferracioli that is not already there for other accounts. In one sense, I agree: other views may fare better or worse depending on which metaphysical view of pregnancy is true. But in another sense, I disagree: If the moral commitment account is supposed to ground moral parenthood in every case, and the moral commitment account has strict motivational requirements, then it is a unique problem for the moral commitment account that there is no apparent view of pregnancy that could make Ann morally committed while she seems to fail those requirements. This is not as much of a problem for other views that lack the moral commitment account's strong motivational requirements. For example, the investment theory of moral parenthood says someone's claim to moral parenthood is grounded in their work toward the child's development. See Millum, *The Moral Foundations of Parenthood*, 21. On this theory, Ann still counts as the moral parent despite her ambivalence because of the gestational work she has put into the development of her fetus, even if she is ambivalent about the fetus. Or the causal theory of parenthood says the relevant cause of a child's existence has the *prima facie* right to claim the right to raise that child. See, e.g., Archard, "The Obligations and Responsibilities of Parenthood." Whatever demerits this view has—and see Ferracioli, *Parenting and the Goods of Childhood*, chs. 2.2–3 for a convincing discussion of them—it has no problem explaining how Ann is the moral parent of the fetus: she is the relevant cause of its existence. Or there is the project view, whereby procreators have a right to continue their procreative projects so long as their doing so does not violate the rights of others. See, e.g., Richards, *The Ethics of Parenthood*, ch. 1; and Lange, "A Project View of the Right to Parent." Ann may have a right to continue her parental project regardless of how she feels about the project. So while it is true that pregnancy is a problem all views must deal with, and work on the metaphysics of pregnancy will help adjudicate that problem, the strong motivational requirements of the moral commitment account make ambivalent pregnancies a unique problem for the moral commitment account that other views do not face.

On an overly simplistic understanding of the moral commitment account, one might think Billy's fans are his moral parents because they seem morally committed to him. But Ferracioli notes that merely committing to a child does not guarantee becoming that child's moral parent when that child already has morally committed parents:

> My moral commitment to my child is violated [if someone] unilaterally inserts himself into my family life. This is because for any additional person who commits to my child without my consent, my ability to engage in actions that adequately express recognition of the value of my relationship with the child is severely compromised. Implications may include, for example, the inability to see my child as much as it is good for our relationship and the inability to forbid her to engage in what I take to be risky activities. (50)

Ferracioli's response explains nicely why Billy's thousands of adult fans should not count as moral parents simply by attempting to morally commit to him. If they become Billy's parents, their involvement in Billy's life would essentially eradicate Billy's original parents' abilities to express moral commitment to him. Moreover, Billy is psychologically incapable of having relationships with hundreds of parents, so this is not in Billy's interest either.

However, there are other cases in which a third party would either very minimally restrict an original parent's moral commitment or even *enhance* their ability to express themselves morally. Here is another case:

> *The Kindergarten Teacher*: Nikhil is a single father raising his precocious five-year-old, Jimmy. One day, Jimmy's kindergarten teacher, Lisa, notices Jimmy is an excellent poet and forms the desire to raise Jimmy. Nikhil does not want to co-parent Jimmy with Lisa.

The only other person who wants to raise Jimmy is Lisa. It would not take much to accommodate a custodial schedule between two adults—co-parents do this all the time. This arrangement might even *increase* Nikhil's ability to morally commit to his child. After all, single parenting is very hard; a parent in this situation must split their time because they are working on one salary to support their household.[12] Even if Nikhil dislikes Lisa, her involvement in Jimmy's life—whether through additional income, free childcare, etc.—would enhance Nikhil's life by reducing the socioeconomic strains of being a single parent. So, not only would Lisa's involvement minimally conflict with Nikhil's plans, but there is good reason to think it would enhance Nikhil's relationship with his child.

12 I thank Matthew Lee Anderson for suggesting this point to me.

Ferracioli could respond in the following way: Lisa is not morally committed to Jimmy because she is in "violation of a basic moral requirement not to significantly jeopardize the moral commitments of others" (50), particularly Nikhil's commitment to raising Jimmy *as a single parent*.[13] So even if Lisa's involvement would reduce Nikhil's troubles as a single parent, her involvement undermines Nikhil's moral commitment because he wants to raise Jimmy alone.

However, moral commitments generally do not have this sort of exclusionary power.[14] Recall Ferracioli's paradigmatic example of a moral commitment: humanitarian aid. Professionals who commit themselves to this endeavor count as morally committed "because saving the lives of innocent people is an exceptionally valuable activity to engage in but also because it is an activity pursued without recourse to gross human rights violation" (39). But nothing about this moral commitment grounds an exclusionary right against others who want to join this cause. This would be the case even if the professionals initially set out to help others on their own and even if the involvement of others would reduce the professionals' abilities to express themselves to those in need. Similarly, Nikhil's intention to raise Jimmy on his own does not on its own ground an exclusionary right against Lisa.

The moral commitment account is supposed to be an account that grounds the moral parent's exclusionary right to keep other adults from parenting their child without their consent. I am not disputing that an adequate account of moral parenthood should include this normative power. I *am* disputing that the moral commitment account can explain how moral parents have this power against people like Lisa. If Ferracioli wants morally committed parents to have this power, she must clarify how a moral commitment to a child has an exclusionary power that most other moral commitments lack.

## 4. CONCLUSION

Ferracioli provides the ethics of parenthood literature with a novel account of moral parenthood. However, several cases show the moral commitment account is either not truly monistic or too weak to ground the sort of exclusionary power we typically associate with moral parenthood.

*Baylor University*
*nicholas_hadsell1@baylor.edu*

13 I thank an anonymous referee for raising this point.

14 I thank an anonymous referee for suggesting the following counterexample.

## REFERENCES

Archard, David. "The Obligations and Responsibilities of Parenthood." In *Procreation and Parenthood*, edited by David Archard and David Benatar. Oxford University Press, 2010.

Bayne, Tim, and Avery Kolers. "Toward a Pluralistic Account of Parenthood." *Bioethics* 17, no. 3 (2003): 221–42.

Ferracioli, Luara. *Parenting and the Goods of Childhood.* Oxford University Press, 2023.

Lange, Benjamin. "A Project View of the Right to Parent." *Journal of Applied Philosophy* 1 (2023): 1–23.

Millum, Joseph. *The Moral Foundations of Parenthood.* Oxford University Press, 2017.

Richards, Norvin. *The Ethics of Parenthood*. Oxford University Press, 2010.

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
*http://www.jesp.org*
ISSN 1559-3061

The *Journal of Ethics and Social Philosophy* (JESP) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge. Articles are typically published under the CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 license, though authors can request a different Creative Commons license if one is required for funding purposes.



Funding for the journal has been made possible through the generous commitment of the Division of Arts and Humanities at New York University Abu Dhabi.

جامعة نيويورك أبوظبي
NYU ABU DHABI