RETRIBUTIVE CAUTION, GUILTWORTHINESS, AND THE RATIONALITY OF ANGER

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy
by
Austin Paul Duggan
August 2017

RETRIBUTIVE CAUTION, GUILTWORTHINESS, AND THE RATIONALITY OF ANGER

Austin Paul Duggan, Ph.D.
Cornell University 2017

Abstract

On a basic desert conception of moral responsibility, any agent who freely, knowingly, and inexcusably acts immorally deserves a negative response *just* because s/he so acted. The predominant view is that blame is the negative response at issue. I argue that, even if we know that an agent has freely, knowingly, and inexcusably acted immorally, there is no reliable evidence that this fact alone renders the agent deserving of blame or other form of censure. We ought, then, to be skeptical about the desert of blame, and in turn forego blaming or censuring on that basis. I argue that this does *not* entail that agents are never morally responsible for immoral acts. Some agents may still deserve to feel guilt for their immoral acts, and it is this guiltworthiness that explains their moral responsibility. Moreover, I argue that skepticism about the desert of blame does not impugn the rationality of our blaming emotions. These views stand in stark contrast to extant views concerning blame and moral responsibility, and offer solutions to heretofore unresolved issues in ethics.

Biographical Sketch

Austin received a B.S. in Mass Communication from Middle Tennessee State University in 2003. From 2003 to 2010 he worked as a musician and sound engineer in NYC. In 2010, he received a B.A. in Philosophy from CUNY, Brooklyn College, and was graciously accepted into the Ph.D program in the Sage School of Philosophy at Cornell. He is a first-generation college student.

To Dad, for obvious reasons,
and to Amy, a fellow lone wolf.

Acknowledgments

This project began with a few questions about the sort of 'basic' desert Derk Pereboom cites throughout his work on free will. Much of it now consists of responses to or developments of his ideas on the subject. I've benefitted tremendously having him as my advisor. He's been generous with his time, encouraging with his comments, and inspiring in his work ethic. With luck, his influence on me will endure.

I'm also indebted to Kate Manne, Michael McKenna, Michelle Kosch, and Nick Sturgeon. Kate provided insightful commentary on my work, particularly in the early stages. She was especially helpful on the first drafts of what eventually became "Moral Responsibility as Guiltworthiness". Michael meticulously combed through later versions of that paper, and invited me to present it at the University of Arizona. These contributions were essential to the paper's development, as well as my own. Michelle offered excellent publication advice for the papers in this collection. She also revealed tensions between them that I would never have discovered otherwise. Nick's guidance during my coursework influenced my thinking on most ethical issues, and is a major reason for my shift to ethics in graduate school. His erudite comments during my defense compelled me to make revisions to sections of argument I had thought unproblematic.

Numerous graduate students provided helpful feedback in my workshops at Cornell. Of particular value were the contributions of Katie Mathie-Smith, Philippe Lemoine, Kim Brewer, Fran Fairbairn, David Kovacs, Lucia Munguia, Ian Hensley, Alicia Patterson, and Quitterie Gounot. I also owe a good deal to Justin Steinberg, Andrew Arlig, and Christine Vitrano, the CUNY professors who set a good example for me early on, and to Brenna Burlingame for aiding me in ways that no one else could. Nancy Brown, in a number of leisurely conversations with me about this dissertation, shared many of my quirky intuitions about desert. It helped.

Finally, much of this dissertation was written at Gimme Coffee (Ithaca, NY), Variety Coffee (Brooklyn, NY), Crema (Nashville, TN), and Woodcat (Los Angeles, CA). Special thanks are due to the baristas and owners for providing comfortable workspaces, and for enduring my persnickety coffee preferences.

Table of Contents

Introduction

On a *basic desert conception of moral responsibility* (henceforth simply 'moral responsibility'), any agent who freely, knowingly, and inexcusably acts immorally deserves a negative response *just* because s/he so acted. Historically, the dominant focus in explorations of this concept has been on its *agential* features; the kind of free will and moral sensitivity required to be morally responsible, whether or not anyone has these capacities, and how we ought to treat people if they lack them. Since Peter Strawson's influential paper *Freedom and Resentment*[1], work on moral responsibility has broadened to also include a focus on the relation between moral responsibility and the *interpersonal* nature and norms of our blaming practices.

I suspect that it is largely due to this Strawsonian emphasis that moral responsibility has come to be seen as essentially connected to these blaming practices. The predominant view is that an agent is morally responsible for an immoral action just in case s/he deserves to be blamed for it[2], and that it's irrational to blame those who aren't morally responsible for their immoral actions.[3] Furthermore, due to the continued focus on agential concerns, the question of whether or not a person is morally responsible is still seen as solely a question of whether or not s/he has the requisite agential capacities to be blameworthy. I think each of these views is mistaken. Once we

---

[1] Strawson 1982.

[2] Feinberg 1970, 55-94; Gibbard 1990, 42; Wallace 1994, 76-77; Fischer and Ravizza 1998, 5-8; Strawson, G. 2002, 452; Bennett 2002; Clarke 2005, 21; Fischer 2006, 63; Darwall 2006 and 2009; Shoemaker 2007 and 2011; Waller 2011, 2-5; McKenna 2012, 150; Boxer 2013; Scanlon 2013, 101- 102; Pereboom 2014, 2.

[3] Honderich 1988, 583-585; Smilansky 2002, 501; Scheffler 2003; Strawson, G. 2008; Pereboom 2014, 128-129.

shift the focus away from the interpersonal and agential features of moral responsibility towards its oft-neglected intrapersonal and normative features, we discover that moral responsibility is only contingently related to our blaming practices, and in the end only sometimes justifies them. Or so I argue in this collection of papers. Here's a sketch of the project.

I first set the stage for much of the work that follows by defending a *justifying* conception of the desert that constitutes moral responsibility. On a justifying conception, when an agent 'deserves' a negative response, there is a defeasible non-instrumental justification for responding to the agent in that way. There are rivals to this conception. By 'deserves a response', some - most notably George Sher, Michael Zimmerman, and Ishtayique Haji - mean only that the beliefs partly constitutive of that response (e.g. the beliefs constitutive of the negative reactive attitudes) are justified and true. Alternatively, by 'deserves a response', some - for example Nathan Hanna and (on one reading) Tim Scanlon - mean only that it is less bad that the agent receive that response than it would have been, had s/he not deserved it. I argue that the justifying conception explains paradigm moral responsibility practices, distinguishes moral responsibility from responsibility for imprudent actions, and explains why some individuals seem 'more' deserving of a negative response than others. I argue that, since these features are crucial to our core conception of moral responsibility, rival views are at best revisionist.

With this defense in hand, I turn to the question of whether or not blame can be deserved in the sense at issue in attributions of moral responsibility. Many claim that those who are responsible for wrongdoing deserve to be blamed, and that an agent's blameworthiness at least sometimes permits blaming or otherwise censuring them. I

argue in "The Case for Retributive Caution", that blameworthiness does not offer this permission. I avoid the typical agential concerns and just assume throughout that people sometimes act immorally without excuse and with the requisite free will and moral sensitivity. I argue that, even if we assume that agents freely, knowingly, and inexcusably act immorally, there is currently no plausible theory explaining how these facts alone offer a non-instrumental justification to blame or otherwise censure such agents. Moreover, our blameworthiness intuitions are too likely to be biased by an unconscious process I call 'anger affirmation' to be reliable. I argue that the unreliability of intuition, together with the absence of a plausible theory of blameworthiness, undermines adequate epistemic justification for the belief that anyone is blameworthy for acting immorally. Our concept of blameworthiness, however firmly entrenched, is unlikely to apply. I argue that we ought, then, to practice *retributive caution* by refraining from blaming or otherwise censuring agents on the basis of their alleged blameworthiness.

At first glance, it might seem that this skepticism about blameworthiness commits one to skepticism about moral responsibility for immoral acts. This is because the predominant *blame-focused* view of moral responsibility is that an agent is morally responsible for an immoral act just in case s/he is blameworthy for it. However, I argue that beliefs about deserving to feel moral guilt for immoral acts are on surer epistemic footing, and that we may sometimes have sufficient epistemic justification for such beliefs. I then argue in "Moral Responsibility as Guiltworthiness" for a *guilt-focused view of moral responsibility.* The desert of blame, I argue, does not explain moral responsibility for immoral acts, as the current literature suggests. Rather, an agent is

morally responsible for an immoral act just in case s/he deserves to feel moral guilt for it. I argue that this conclusion explains puzzling features about moral responsibility that rival views cannot. Most notably, it explains why those who commit suberogations, those inured to the intended effects of blame, and those who do not fully understand what Darwall calls 'second-personal' reasons are blameless but nevertheless seem morally responsible for their immoral acts. So, skepticism about blameworthiness does not entail skepticism about moral responsibility.

Many philosophers argue that skepticism about blameworthiness at the very least impugns the rationality of the reactive anger partly constitutive of blame, and therefore generates a strong reason for skeptics such as myself to either relinquish doubt or eschew anger. Appealing to the conclusion Gary Watson draws in his influential discussion about the sadistic murderer Robert Harris, these philosophers argue that it is nearly impossible for any sane adult to target an agent with the anger that is constitutive of blame without also believing that the targeted agent deserves it. I call this view *retributivism about anger*. In my final paper, "Skepticism and the Rationality of Anger", I debunk the Harris explanation, and argue for a view I call the *justifying view of anger.* Roughly, this is the view that an agent can target an individual with anger if she believes she is sufficiently justified in doing so, and values acting on that justification more than she disvalues harming the individual. I then argue that there are non-desert based justifications that satisfy this condition. I cite as paradigms two kinds of cases: cases of moral education in which we value the reform of the immoral agent and moral community more than we disvalue the temporary harm anger causes, and cases of moral protest in which we react angrily to an immoral but nevertheless undeserving

actor in order to overcome or take a stand against the immoral system of which they are a part. Given our values, lacking the belief that such individuals deserve to be targeted by anger might decrease the severity of our anger, but it surely cannot extinguish it. So although skepticism about blameworthiness undermines a common justification for the reactive anger constitutive of blame, it does not impugn the rationality of blaming agents in reaction to their immoral acts.

In short, my view is that we ought to be skeptical about blameworthiness, but that this does not have the drastic implications for moral responsibility that the current literature suggests, since moral responsibility for an immoral act is neither explained by blameworthiness, nor presupposed by blame.

Defending a justifying conception of desert

It is generally agreed that desert is a triadic relation between a deserving subject, an object that the subject deserves, and a basis for the subject's deservingness of the object.[4] Aside from that feature, it is unlikely that there can be any non-trivial analysis that applies to all desert claims.[5] I leave it an open question whether or not there can be. Instead, and as indicated in the introduction, I limit myself in this work to the kind of desert that an agent generates *solely* by virtue of acting in a certain way with the requisite agential capacities. In the ethics and political philosophy literature this sort of desert is known as 'raw'[6], or 'pre-institutional'[7] desert. In the philosophy of agency literature, it is known as 'basic'[8] desert. Though I enjoy the visceral character of the first descriptor, I will in the remainder of this work employ the last.

There are two sorts of desert that are potentially 'basic' in this way.[9] *Distributive desert* is desert of a portion of resources, benefits, or burdens that are, perhaps necessarily so, distributed by institutions. We might say, for example, that Jennifer deserves more of the total pay available to workers for staying overtime, while Jane, her workmate, deserves less because she slept on the job. There is also the desert of

---

[4] This was made explicit in Rescher 1966, 62; Feinberg 1970; and Kleinig 1971, and has been assumed in much of the literature since then. One notable exception is Vilhauer 2009.

[5] Sher 1987, xii.

[6] Kleinig 1971.

[7] Scanlon 1988; Rawls 1991; Scheffler 1992; McLeod 1999; Moriarty 2002.

[8] Pereboom 2001, xx; 2007, 170. Pereboom's 'basic' title is adopted by McKenna 2012, 121, and Shabo 2012b, 159. Untitled characterizations of this sort of desert are prevalent in the free will literature. See, for example, G. Strawson 2002, 452-455; Waller 2011, 2-5; Clark 2005, 21.

[9] For a concise overview see Sher 1987, 6-8.

negative responses such as blame, guilt, or punishment generated solely by those actions that are immoral, and done freely, knowingly, and inexcusably.[10] It is only this latter sort of desert that concerns me in this work.

What is it to 'deserve' some negative response for an immoral action? I join the near consensus that desert constitutes a defeasible normative justification in the form of the 'good' or the 'right'. More specifically, I endorse

> *The justifying conception of desert (JD)*: When an agent *A* deserves a negative response *R* for an immoral action, there is some non-instrumental good in seeing to it that *A* receive *R*[11], or there is a defeasible non-instrumental reason to see to it that *A* receive *R*.[12]

There are two rival conceptions to *JD*. In what follows, I defend *JD* against arguments in favor of each rival conception.

2. Against the Epistemic Sense of 'Deserves'

Michael Zimmerman claims that there is an epistemic use of the term 'deserves' such that, by '*A* deserves a negative response *R*', some mean only that there is a reason for agents to hold the doxastic attitudes involved in that response, where the

---

[10] 'Moral' desert is sometimes used to specify this kind of basic desert. See, for example, Rawls 1991, 2003; Scheffler 2000; Shafer-Landau 2000; Miller 1999, 134, 148.

[11] Ross 1930, 134-138; Sher 1987, 194-198; Hurka 2001, 8-12; Moriarty 2003, 520; Bennett 2002, 147; Kagan 2003, 93; McKenna 2012, 172; Pereboom 2014, 137.

[12] Mundle 1954, 217; Feinberg 1970, 60; Murphy 1971 and 1973; Kleinig 1973, 62-3; Zimmerman 1988, 162; Moore 1993, 15; Schmitdz 2002, 774.

reason is generated solely by the truth of, or evidence for, those doxastic attitudes.[13] Philosophers sometimes employ this use of the term in theories of 'appraisability' or 'attributability', which concern the justification for holding certain sorts of immoral appraisals of agents, or attributing agents with certain immoral characteristics.[14] George Sher appeals to this sense of desert, claiming that

> We must take 'X deserves blame' to mean no more than that blame directed at X is justified or appropriate[15], [where] the norms that render it appropriate to have the desire-belief combinations that I have said add up to blame stand revealed as those that require that we believe propositions that are true and that we accept moral principles that are justified.[16]

Call this the *epistemic conception of desert*, or *ED. ED* is too weak to accommodate two core intuitions about desert. The first core intuition concerns a constraint on the relation between affect and deserved response. Some responses affect the deserving agent in some way. That is, they ensure the deserving agent a particular physical or mental state, behavioral response, or level of well-being that would not have occurred otherwise. For example, an agent might be affected by expressions of blame in that she

---

[13] Zimmerman 1988, 152.

[14] For these uses of 'desert', see Zimmerman 1988, 152; Haji 1998, Sher 2006, 93-114, and Boxer 2013, 38. For attributability more generally, see Watson 1996; Arpaly 2006, 9-39; Shoemaker 2011; Nelkin 2011, 34-35.

[15] Sher 2006, 86.

[16] Sher 2006, 130.

is caused to feel bad, to scoff, to believe herself blameworthy, to decide to be indifferent, to apologize, to be worse off, or simply caused to hear an agent blaming her.

By 'blame', Sher is merely appealing to a set of doxastic attitudes. This attributive sense of blame[17] does not necessarily dispose the blamer to act. This is because the attitudes that constitute such blame are all private, unemotional, and need never result in any outward change in the blamer. So such a response does not affect the targeted agent. As Sher points out, "even when someone receives his full measure of deserved blame, his receiving it need not affect his life at all."[18] This does not comport well with conclusions about other kinds of cases. As Sher himself points out, "in many other contexts a person's getting what he deserves necessarily does have some impact on his life."[19]

Suppose Jenny doesn't practice as hard for her competition as her competitors. As a result, she comes in last. Intuitively, Jenny deserves her loss. Suppose Sadam's wanton drinking is hurting his health. He solemnly promises himself he'll quit. But he doesn't. So he feels guilty. Intuitively, Sadam deserves to feel guilty. In each of these cases, the object of desert affects the deserving agent, and needn't affect anyone else. Each deserving agent's desert is theirs and theirs alone. This is evidence, as both John Kleinig and Joel Feinberg note in their work on the issue, that an agent's receipt of her desert must affect her.[20] "Deserved treatment", says Kleinig, "is not something toward

---

[17] See my "Moral Responsibility as Guiltworthiness" for further details.

[18] Sher 2006, 86.

[19] Sher 2006, 86.

[20] Feinberg 1970, 61; Kleinig 1971, 72.

which we remain indifferent."[21] Sher recognizes that such a constraint is at play in other sorts of cases, conceding that

> At first glance, the norms may not seem to explain [deserved blame], since what they demand is...only that other people (or the wrongdoers and bad people themselves in another capacity) have certain beliefs and desires whose propositional objects refer essentially to them. Because the norms are addressed to potential blamers rather than to potential blamees, their demands may seem ill-suited to support a reconstruction of what the latter agents deserve.[22]

Sher argues that the violation of the constraint that desert is affective is acceptable in cases of desert for immoral actions because it is a consequence of accepting (i) that blame is the object of desert for immoral actions,[23] and (ii) that blame is merely some combination of dispassionate doxastic attitudes.[24] I argue in "Moral Responsibility as Guiltworthiness" against (i). But even if we grant it, (ii) is not firmly established. And there are strong intuitions against it. Consider a paradigm example of a wrongdoing. Nathan Hanna has us imagine that "a general orders the kidnapping, brutal torture, and execution of innocent dissidents."[25] Suppose she does this freely, and knowing full well

---

[21] Kleinig 1971, 72.

[22] Sher 2006, 130.

[23] Sher 2006, 70, 80-81.

[24] Sher 2006, 86.

[25] Hanna 2012, 46.

that it is inexcusably wrong. Now consider a case in which the General is targeted with the sort of aretaic blame to which Sher appeals.

> The General's suspicious lieutenant Alfredo discovers her wrongdoing. Alfredo concludes that she is blameworthy, becomes angry at her, and desires that she not have committed the wrongdoing. In a sense, Alfredo 'blames' the General. Although this blame constitutes a change in Alfredo's attitudes, it does not at all alter his behavior. Alfredo never communicates his emotions or attitudes to the General.

Alfredo 'blames' the General in the aretaic sense. Alfredo is certainly epistemically justified in holding those attitudes involved with his blame. So on an epistemic conception of desert, the General gets all that she 'deserves'. But she clearly does *not* get all that she deserves. The object of the General's desert must affect her in some way. Desert is affective. So *ED* is false.

Sher's claim that (i) blame is the object of desert for morally bad actions or characteristics has some initial plausibility because we often allege that immoral agents deserve blame. However, we have in mind not just the collection of attitudes that are sometimes called 'blame', but rather the passionate expression of those attitudes. As I detail in the next two works in this dissertation, 'overt' blame may include many, perhaps all, of the features of attributive blame. But it also includes an overt expression of anger. (i) is only plausible when it refers to overt blame. But (ii) refers to blame that is not overt in this way. So Sher's two-step argument in support of *ED* is unsound. To deserve a

negative response is to deserve something affective, and not merely the non-affective attitudes involved in that response. That is why most philosophers, including Sher in an earlier work[26], offer conceptions of desert that provide a *normative* rather than an epistemic justification. Alfredo, for example, is not merely justified in appraising the General negatively. All else being equal, the General's desert also justifies Alfredo in affecting the General harmfully by way of overt blame or even punishment. *ED* cannot offer such a justification. *JD* can.

Some may protest that this dispute is merely verbal. There is, they might say, an epistemic sense of 'deserves', and a normative sense of 'deserves'.[27] The former justifies negative responses that are not affective. The latter justifies negative responses that are. I'm happy to concede the point. However, I think that the epistemic sense of 'deserves' is, then, nothing more than a term of art that can do no useful theoretic work. After all, 'desert' is a misleading title for a justification for merely believing propositions that are true and accepting moral principles that are justified. We do not typically say that agents like the General 'deserve' others to correctly appraise her. Indeed, almost all of those working on blame do not speak of desert when referencing such a justification, but instead of 'warrant', 'aptness', 'fittingness', or, more directly, of 'epistemic justification'. Linguistic practices are sometimes misguided. But surely the fact that an epistemic sense of deserves is not what people mean to pick out when they use the term is, in the absence of a compelling argument to the contrary, sufficient evidence that the epistemic sense of 'deserves' is not at play.

---

[26] Sher 1989, xi.

[27] Zimmerman 1988.

3. Against Weakened Normative Conceptions of Desert

In opposition to *ED*, I think we should endorse *JD*. As explained above, the negative responses that immoral agents intuitively deserve - blame, guilt, punishment, and the like - are all affective. They are also typically harmful. That troubles some philosophers. Tim Scanlon has consistently held that "it is never a good thing, morally speaking, for anyone to suffer, no matter what they have done."[28] His remarks indicate the same stance on reasons.[29] He rejects *JD* on that basis. However, the conclusion that those who freely, knowingly, and inexcusably act immorally do *not* deserve a harmful response is extremely counterintuitive. So those like Scanlon who reject *JD* typically don't wish to renounce desert altogether. Rather, they aim for a weaker sense of desert that can be supplemented by further justifications. Scanlon has recently endorsed such a view, claiming that an agent's desert of blame can weaken our reasons to offer the agent good will, friendliness, sympathy, and help.[30] Nathan Hanna[31], developing remarks by Fred Feldman[32], has also offered such a view. On Hanna's view,

> *The discounting conception of desert (DD)*: When *A* deserves a negative response *R* for doing wrong, her receiving *R* is not as bad as it would be if she didn't deserve it, or there is less reason to prevent her from receiving *R*. But her

[28] Scanlon 2013, 104. Also see Scanlon 1998, 274.

[29] Scanlon 2013, 107-108.

[30] Scanlon 2013, 105-106.

[31] Hanna 2012.

[32] Feldman 1995.

receiving *R* is still intrinsically bad. And there are no reasons to see to it that she receives *R*.

In what follows, I argue that *DD* (and, *a fortiori*, *ED*) is too weak to accommodate core intuitions about desert that *JD* can, and that there are no good arguments favoring *DD* over *JD*. So *JD* is preferable to *DD*.

Reconsider the case of the General. Suppose that, in response, Alfredo harms the General by way of an expression of overt blame or minor punishment. Given moral prohibitions against harming people for no good reason, Alfredo needs a viable justification for his harmful response. All else being equal, the fact that the General deserved a harmful response to her wrongdoing seems to constitute such a justification. But the mere fact that it is less bad to respond harmfully to the General, and there is less reason to prevent her from being harmed, cannot alone offer a viable justification for harming her. So the *DD* interpretation of desert cannot accommodate our intuitions here. *JD* can.

Alternatively, suppose that those authorities in the General's community who have the relevant moral standing to punish wrongdoers learn of the General's wrongdoing. Suppose further that these authorities purposely neglect to in any way respond to the General for her wrongdoing. Intuitively, the General's desert of a harmful response all by itself renders their failure morally deficient. But the mere fact that it is less bad to respond harmfully to the General, and there is less reason to prevent her from being harmed, cannot explain this. *JD* can. Such intuitions are persistent and commonplace enough that rejecting them in favor of *DD* would seem to require a

substantial revision to our moral thought.[33] So the onus is on the advocates of *DD* to offer compelling arguments in favor of their view.

One argument against *JD* is that it is morally indefensible, given the strength of general moral prohibitions against harming others.[34] However, the claim that there are moral prohibitions against harming others "no matter what they've done" is precisely what's in question. Many people's intuitions counsel against that conclusion. To reject those intuitions on the basis of the alleged obligation doesn't bode well for the *DD* advocate. After all, any moral theory that prohibits harming others will almost certainly obligate us to prevent the harming of others. Intuitions in favor of *DD* are contrary to this latter obligation. So on this strategy, *DD* intuitions would also have to be rejected.

Another kind of argument against *JD* concerns its questionable epistemic origins.[35] I defend such an argument in "The Case for Retributive Caution". But in this case employing such an argument would yield another pyrrhic victory for *DD* advocates. The intuition that an agent's wrongdoing provides an objective reason, or makes it non-instrumentally good, to respond to her harmfully almost certainly shares its epistemic origins with the intuition that an agent's wrongdoing weakens reasons against responding to her harmfully, or makes it less bad to respond to her harmfully. So if the epistemic origins of *JD* judgments are suspicious, so too are those of *DD* judgments.

---

[33] R.J. Wallace concedes that weakened conceptions of desert like *DD* may be revisionist for precisely this reason (1994, 228).

[34] Wallace 1994, 226-227; Scanlon 2008, 189.

[35] Murphy 2007.

A much better strategy for the *DD* advocate is to try to show that our intuitions about desert themselves support *DD* rather than *JD*. Nathan Hanna offers two such arguments. The first involves a case of severely imprudent behavior. Suppose that

> Johnny, a carefree undergrad, goes to a frat party. He's eager to impress his bros and after a couple drinks he hits upon a daring scheme. He clambers to the roof of the frat house with a sheet, bellows out his favorite beer's name while striking a macho pose, hurls himself off the roof, and tries to parachute onto the lawn, aiming for a convenient spot near the keg (living dangerously is thirsty business). Unsurprisingly, he breaks his ankle.[36]

Hanna contends that

> Dj: Johnny deserved to break his ankle.

But Hanna claims that it's also intuitive that "It's intrinsically bad that Johnny broke his ankle. And there was no reason - not even a weak one - to bring this about for its own sake."[37] Hence, *JD* is false in this case. But even those who have this intuition will likely agree that Johnny's ankle-breaking is *less bad* than it would have been, had Johnny suffered it through no fault of his own, a conclusion that supports *DD*.

---

[36] Hanna 2012, 43.

[37] Hanna 2012, 43.

Dj is intuitive. And Hanna's analysis is compelling. I take it to show that *JD* cannot be the relation at stake in cases like Johnny's. But there are different kinds of desert. And the sort that is at stake in cases like Johnny's is conceptually distinct from the sort of 'moral' and 'basic' desert that is at stake in cases like the General's. So the Johnny example does not impugn *JD* as best capturing our intuitions about a kind of desert that many take to be at the core of our judgments concerning moral responsibility, as well as justified punishment and blame.

Consider first my claim that there are different kinds of desert. It is commonplace to hear that a person deserves a prize, deserves to win the lottery, deserves her lot in life, deserves better weather, deserves punishment, or deserves to be better informed. The first pages of a Google news search just now (try it yourself!) reveals the claims "Plato deserves better than a stint at the Googleplex", and "everyone deserves a place at the library." As indicated above, these desert claims are too diverse to be given a unifying analysis that is non-trivial.[38] The way in which a person might deserve better weather, for instance, seems different in kind from the way in which a person may deserve punishment, and these in turn seem different in kind from the way in which everyone might deserve his or her place at the library.

Consider next my claim that the sort of desert at issue in cases like Johnny's is distinct from the sort of desert at issue in cases of immoral actions like the General's. There are three considerations supporting this conclusion. First, if Johnny deserves his ankle-breaking, he deserves it because it is a predictable consequence of his action.[39]

---

[38] Sher 1987, xii.

[39] Sher 1987, 41.

But the General's desert of something harmful has *nothing* to do with whether it is a predictable consequence of her immoral action.

Secondly, the sort of desert at issue in the General case clearly constitutes her moral responsibility. This makes sense in the General case, considering that the General deserves by virtue of immoral actions. But there is nothing necessarily morally significant about Johnny's imprudence. His imprudence doesn't constitute anything wrong or morally bad. And this is why, although he may be responsible for his imprudent action, it seems incorrect to say that he is *morally* responsible for it.

Lastly, the sort of desert in the General case is a sort of desert that is *not* in virtue of consequentialist considerations. But it seems to me that desert of the harm that results from imprudence requires that the harm help the imprudent in some way. For example, Johnny seems to 'deserve' to break his ankle for pulling such a dangerous stunt because he's likely to 'learn his lesson' from such harm, or at the very least set an example for others. If the event does nothing to dissuade Johnny or others from further dangerous stunts, then what happens to Johnny seems not deserved but tragic. These considerations support the conclusion that *DD* applies to a sort of non-moral desert that is conceptually distinct from the sort of 'moral' and 'basic' desert at issue in the General case.

Anticipating such a response, Hanna offers a second argument against *JD* involving the General case. Recall that the General orders the kidnapping, brutal torture, and execution of innocent dissidents. Suppose that, in retaliation for the General's

actions, rebels kidnap her, brutally torture her, and execute her.[40] Hanna claims that, when this happens,

G: The General gets what she deserves.

Consider those who accept *G*. Together with *JD*, *G* entails that there is some non-instrumental good, or there is a defeasible objective reason to see to it, that the General is kidnapped, tortured, and executed. Call the acceptance of both *G* and *JD* the 'strong view'. Hanna contends that the strong view cannot be right. "That such treatment can never be intrinsically good and that we can never have a reason - not even a weak one - to inflict it for its own sake is compelling. Some things just seem off limits in this sense."[41] But it still seems that the General's desert, "weakened the reasons not to treat her that way - so that what the rebels did to her wasn't as wrong as what she did to any of her victims - and [her desert] weakened the reasons to do things like help her and sympathize with her."[42] Hence we should accept *DD* to explain *G*.

However, those endorsing the strong view can consistently claim that some harmful responses to wrongdoing are 'off limits'. Intuitively, some humans will have the moral standing to inflict some *less* harmful response on the General; perhaps overt blame or imprisonment. But it is plausible that the epistemic limitations, emotional integrity, and moral imperfections of human persons preclude any actual human person from having the moral standing to kidnap, torture, and execute another. Perhaps only a

[40] Hanna 2012, 46.

[41] Hanna 2012, 46-47.

[42] Hanna 2012, 47.

God could have such moral standing. So the strong view that there is some non-instrumental good, or there is a defeasible objective reason to kidnap, torture, and execute the General is consistent with the claim that, since no one has the relevant moral standing to mete out such treatment, such treatment is always 'off limits'.

One response to this sort of position is that a defeasible justification that is always defeated is no justification at all. So some treatment can't be both deserved but universally 'off limits' in the way sketched above. But notice that a defeated desert-based justification to kidnap, torture, and execute the General may still offer sufficient justification to mete out some less harmful response on her. Indeed, I think it is intuitive that, if the General deserves to be kidnapped, tortured, and executed, then, all else being equal, those with the relevant moral standing are justified in meting out some less harmful response upon her. They can blame, punish, or guilt the General, all the while claiming that they are justified in doing so because she deserved much worse. *DD* cannot offer such a justification. *JD* can.

Of course, many will be prone to reject *G* in the first place on the basis of the 'off limits' intuition that there is a limit on the harmfulness of the responses that can be deserved. Most of us are prone to reject *G* in favor of

> *L*: The General deserves something less harmful than the 'off limits' response of kidnapping, torture, and execution.[43]

Hanna claims that *JD* can't accommodate our intuitions about *L*.

---

[43] Hanna 2012, 47.

> Consider any proposed limit on [the harmfulness of deserved responses] and people who reach it. Take two such people who are [harmed] equally and in excess of it, say the General and someone even worse: the Generalissimo. It seems like what happens to the Generalissimo isn't as intrinsically bad. And it seems like there's less reason to do things like help him and sympathize with him. If [*JD* is true] and if there's a limit on [the harmfulness of deserved responses], then we can't appeal to desert to explain important differences between such cases. [But] desert is an obvious place to look for an explanation.[44]

Hanna is surely correct that targeting the Generalissimo with a harmful response is less bad than targeting the General with that same response. He's also correct that there is less reason to help and sympathize with the Generalissimo. Hanna's argument rests, then, on his further claim that, if there's a limit on the harmfulness of deserved responses, then *JD* cannot explain why the harm that befalls the Generalissimo is less bad than the harm that befalls the General, or why there is less reason to help and sympathize with him. But that's false; the justifying conception of desert can explain this fact.

Suppose that the General and the Generalissimo each deserve the most harmful response that can be deserved. Suppose now that we can only respond to *one* of them. Who do we have most reason to respond to? Who is it better to respond to? Clearly, it is

---

[44] Hanna 2012, 47.

the Generalissimo. This is because, assuming both the General and generalissimo deserve the same harmful response, the Generalissimo, being guilty of a greater transgression, is *more* deserving of it than the General.[45] Even if there is a limit on the harmfulness of the response that a person can deserve, there may be no limit on *how much* a person can deserve it. More specifically, *JD* does not preclude the possibility that

> *JDg*: There is some non-instrumental good to degree *i*, or there is a defeasible objective reason of strength *n*, to see to it that the General receives the most harmful response that can be deserved.

> *JDo*: There is some non-instrumental good to a degree greater than *i*, or there is a defeasible objective reason of a strength greater than *n*, to see to it that the Generalissimo receives the most harmful response that can be deserved.

JDg and *JDo* together offer an excellent explanation of why any harmful response that befalls the Generalissimo is not as bad as the same harmful response that befalls the General, and why there is less reason to help him. The Generalissimo is simply *more* deserving of that response, a notion that *JD* can accommodate. So *L* doesn't generate the problem for *JD* that Hanna alleges. So his argument does not show that our intuitions about desert support *DD* rather than *JD*. On the contrary, *DD* requires a

---

[45] See Kagan 2003, 93.

substantial revision to our intuitions about desert. Advocates of *DD* have not shown that we should make such a revision.

This conclusion does not, nor is it meant to, show that anyone ever deserves anything for their immoral actions. Perhaps no one ever does.[46] But insomuch as anyone deserves anything for an immoral action, the justificatory conception of desert is preferable to the discounting and epistemic conceptions of desert. Our conception of desert doesn't merely discount or provide doxastic warrant. It normatively justifies.

[46] Pereboom 2014.

The Case for Retributive Caution

There are senses of free will and moral knowledge that are so stringent in their requirements as to be elusive in a world like ours. But let's assume for the moment that some of us, sometimes, freely and knowingly do wrong without a good excuse. Call those who do *responsible* for their wrongdoing. It's often alleged that such individuals are *blameworthy in a basic-desert sense*. A wrongdoer is blameworthy in this sense (henceforth simply 'blameworthy') just in case the wrongdoer's responsibility alone offers a defeasible non-instrumental justification[47] to censure them in some way; overtly blame them, shun them, punish them, etc.[48] No further facts need obtain for blameworthiness, for example those offered by consequentialist or contractualist theories.[49] Simply, those responsible for wrongdoing deserve to be censured *just* because they are responsible for wrongdoing.

The concept is prevalent. Ross believed it self-evident that the morally vicious deserve to suffer pain.[50] The free will debate largely concerns whether or not causation

---

[47] Some explicitly ground the justification in the good (Ross 1930, 134-138; Sher 1987, 194-198; Miller 1999, 136; Hurka 2001, 8-12; Bennett 2002, 147; Moriarty 2003, 520; Kagan 2003, 93; McKenna 2012, 172; Pereboom 2014; Rosen 2015). Others explicitly ground it in the right (Mundle 1954, 217; Murphy 1971 and 1973; Kleinig 1973, 62-3; Zimmerman 1988, 162; Moore 1993; 15; McLeod 1999, 193; Schmitdz 2002, 774).

[48] Ross 1930, 134-138; Mundle 1954, 217; Feinberg 1970, 55-94; Murphy 1971 and 1973; Kleinig 1973, 62-3; Sher 1987, 194-198; Gibbard 1990, 42; Wallace 1994, 76-77; Zimmerman 1988, 162; Fischer and Ravizza 1998, 5-8; Miller 1999, 136; Hurka 2001, 8-12; Strawson, G. 2002, 452; Bennett 2002; Moore 1993; 15; McLeod 1999, 193; Schmitdz 2002, 774; Moriarty 2003, 520; Kagan 2003, 93; Clarke 2005, 21; Fischer 2006, 63; Darwall 2006 and 2009; Shoemaker 2007 and 2011; Waller 2011, 2-5; McKenna 2012, 150; Boxer 2013; Pereboom 2014, 2; Rosen 2015.

[49] On this point see especially Feinberg 1970; McKenna 2012; Pereboom 2014.

[50] Ross 1930, 136-137.

allows the sort of free will requisite for blameworthiness. Even Scanlon, a former skeptic about desert, has recently conceded that responsible wrongdoers deserve to be blamed.[51] Justice Stewart cites blameworthiness in *Gregg v. Georgia*, the Supreme Court decision that re-animated the death penalty in America, as the primary justification for capital punishment, claiming that the need "to impose upon criminal offenders the punishment they 'deserve'" is a viable justification distinct from deterrence. Research in social psychology confirms that Stewart's sentiment about the death penalty is widespread[52], and that the folk tend to more generally endorse blameworthiness as a justification for punishment instead of instrumental justifications.[53] Tellingly, that research also shows that those few who explicitly deny blameworthiness in favor of instrumental justifications for punishment are nevertheless far more sensitive to non-instrumental considerations, and still punish wrongdoers even when it is made obvious that doing so is suboptimal.

Clearly, blameworthiness is frequently both thought and treated as if it's among the moral considerations permitting us to censure those responsible for wrongdoing.[54] I think that's an unfortunate mistake. In section 1, I argue that an agent's blameworthiness permits one to censure that agent only if one has adequate epistemic justification to believe that the agent is blameworthy. This requires one to have adequate epistemic justification to believe that

---

[51] Scanlon 2013.

[52] Ellsworth and Gross 1994.

[53] Carlsmith, Darley, and Robinson 2002; Carlsmith 2006; Carlsmith and Darley 2008.

[54] See especially Feinberg 1970, 55-94; Gibbard 1990, 42; Wallace 1994, 76-77; Fischer and Ravizza 1998, 5-8; Bennett 2002; Clarke 2005, 21; Fischer 2006, 63; Darwall 2006 and 2009; Shoemaker 2007 and 2011; McKenna 2012, 150; Boxer 2013; Rosen 2015.

(B)lameworthiness: a wrongdoer's responsibility alone offers a defeasible non-instrumental justification to censure them.

One commonly cited source of justification for this belief is intuition. In section 2, I argue that intuitions in favor of *B* are too likely to be brought about by an unreliable process I call 'anger affirmation' to provide adequate epistemic justification. One could have adequate epistemic justification for *B* if there were a plausible theory that supports it. But I argue in section 3 that there is no such theory. The absence of reliable intuition or a plausible theory undermines any other potential sources of justification, for example expert testimony or peer agreement. It follows that, until we have stronger epistemic justification for *B*, an agent's alleged blameworthiness does not permit censuring that agent. I argue in section 4 that we ought, then, to practice *retributive caution* by refraining from censuring agents on the basis of blameworthiness. I conclude in section 5 by considering some objections, and explaining why judgments about the desert of guilt are on surer epistemic footing than judgments about the desert of censure.

1. An epistemic restriction on the permissibility of harm

I begin with some brief remarks about the sorts of responses that are intuitively deserved by the blameworthy. As an expedient, and in full recognition of the complexity of the issue, I'll say that an agent is *harmed* when s/he is made to be in pain, in a state

that s/he dislikes, deprived of pleasure, or deprived of a state that s/he likes.[55] Those who are averse to the terminology are welcome to substitute the term 'hurt' for 'harm'.[56] We sometimes harm others accidentally, unavoidably, or in response to a moral dilemma. Though such inflictions of harm lack a good moral justification, their moral status is unclear. That's not the case for those harms that one can easily avoid inflicting but one knowingly inflicts anyway. Consider the harm of certain blaming responses.

There are two core senses of blame. One is purely attributive. To attributively blame an agent for an immoral act is to believe that the agent has acted immorally, and to wish for the right moral reasons that the agent had acted differently.[57] Henceforth, I will refer to this attributive sense of blame as *moral disapproval*. The sort of blame that concerns me in this paper has an emotive component. To blame an agent in this sense for an immoral act is to feel some form of anger (such as resentment or indignation) towards that agent in virtue of one's moral disapproval of that agent.[58] Henceforth, I will refer to this sense of blame as simply *blame.*

---

[55] Mill 1863; Griffin 1986; Crisp 2006. It is also plausible that an agent can be harmed *objectively* in that s/he is deprived of something that is intrinsically good independently of any feelings or mental states (Moore 1903). One might also think that well-being is constituted only by the combination of subjective and objective features (Kagan 2009; Parfit 1984, 500-501). I focus on the subjective harms of displeasure and dissatisfaction because those are the harms that primarily concern us when censuring (Kleinig 1971; Feinberg 1970).

[56] Thanks to Kate Manne for this helpful suggestion.

[57] Watson 1996; Arpaly 2006, 9-39; Sher 2006, 93-114; Smith 2007, 476; Scanlon 2008, 122-214; McKenna 2012, 21-29.

[58] Feinberg 1965; Strawson 1982; Gibbard 1992; Wallace 1994, 76-77; Watson 1996, 238; Bennett 2002 and 2008; Smith 2007, 477; McKenna 2012, 21-29; Boxer 2013.

Blame is typically *overt* in that it is expressed in a way that affects the target in some way.[59] Overt blame is almost always harmful. Granted, one might be able to overtly blame in a non-harmful manner (perhaps through clenched teeth). However, it is generally agreed that the emotive component of blame is constituted by dispositions to ostracize[60], elicit guilt[61], and elicit repentance,[62] each of which is likely to be harmful when acted upon, all else being equal. It's difficult to overtly blame without acting on these dispositions. So most every instance of overt blame constitutes or counts as an attempt to affect the target in ways that one knows are likely to harm that agent. In order to avoid the ambiguities plaguing the debate about blame and blameworthiness, I use the term 'censure' to apply to this more limited class of blaming responses. Specifically, a censure is

Censure: An instance of overt blame that one knows is likely to harm the target.

Punishment is a paradigm censure.[63] Most instances of other common blaming responses to perceived wrongdoing - public ridicule, shunning, shaming, and the like[64] -

---

[59] By 'affect', I have only minimal criteria in mind. An agent is affected by an expression of blame when that expression produces in the targeted agent a particular physical or mental state, behavioral response, or level of well-being that would not have occurred otherwise. That a response affect an agent is a requisite for that response to be deserved. On this point see Feinberg 1970, 61; Kleinig 1971, 72.

[60] Feinberg 1984, 37; Bennett 2002, 149-152; McKenna 2012, 136.

[61] Gibbard 1990; Bennett 2002, 152-153; Darwall 2006; Smith 2007, 477; Shoemaker 2007, 91; McKenna 2012, 139-140; Boxer 2013, 125.

[62] Bennett 2002, 151; McKenna 2012, 138-141; Boxer 2013, 125.

[63] Feinberg 1965 and 1971, 61.

[64] Wallace 1994, 54.

also fall under the 'censure' category, even if such responses are not always intended to harm their target.[65]

Given the harmful nature of censures, any moral theory that has a prayer of being true holds that a reason to censure an agent permits one to censure that agent only if it is a *good* reason for one to censure. A criterion commonly determining this status is a minimal epistemic standard we apply quite generally to belief-formation. Consider a case in which you come to believe that you have a reason to censure someone. An *adequate* epistemic justification for this belief is, as Walter Sinnott-Armstrong puts it, a justification strong enough that we "ought to believe it as opposed to denying it or suspending belief."[66] Now suppose you formed your belief only because of petty hatred, biased assumptions, or obviously unreliable testimony. Such a belief is not adequately justified. You ought, then, to suspend it, if not outright deny it. *Prima facie*, your supposed reason to censure is not a 'good' reason for you to censure.[67]

I'm agnostic as to whether or not this *prima facie* consideration wins the day all-things-considered for every kind of alleged justification to censure. My argument here is only meant to support the claim that this minimal epistemic standard applies to blameworthiness. Since this claim is the first premise of my argument for retributive caution, I'll call it (i), though those desiring a more descriptive title are welcome to call it 'the caution principle for blame', or simply 'the caution principle':

---

[65] McKenna 2012, 28-29; Scanlon 2013, 103.

[66] Sinnott-Armstrong 2006, 341.

[67] As Derk Pereboom remarks, "where there is a substantial likelihood that one's justification for harming someone is illegitimate, then harming that person on the basis of that justification could well be morally wrong" (2001, 161).

(i) An agent's alleged blameworthiness permits one to censure that agent only if one has adequate epistemic justification to believe s/he's blameworthy.

Note that only permissibility is at stake in the antecedent of (i). The defeasible normative justification presumed to be offered by blameworthiness is not. The distinction is crucial to keep in mind. I follow Gideon Rosen in recognizing the possibility that a wrongdoer can be blameworthy even if no one has epistemic justification to believe them blameworthy.[68] Indeed, it may be that there is always some value in a responsible wrongdoer suffering a censure. We tend to see some good in even the most misguided censures befalling those responsible for wrongdoing in the way that we see some cosmic justice in their suffering an accidental bump on the head. Similarly, there might be some epistemically inaccessible reason to censure every responsible wrongdoer. Perhaps there is even a sense in which the censuring of a blameworthy agent 'ought' to occur in the way that the weather 'ought' to be nice.[69] But that does not permit those who lack the aforementioned epistemic justification to inflict censures on the blameworthy. There are two reasons why.

The primary reason has to do with *risk*. Censuring is almost always harmful overall. So censuring an agent in response to their alleged blameworthiness when one lacks adequate epistemic justification for the belief that s/he's blameworthy is to take a substantial risk that one is causing overall harm to that agent for no reason. There might be some moral good in taking this risk in a case in which the overall harm is small and

---

[68] Rosen 2015, 69.

[69] Pereboom 2014, 139-140.

the alleged reason for censure is a consequentialist one, for example mildly censuring an agent in order to prevent the small possibility of catastrophic harm. But taking a substantial risk that one is harming an agent in response to a non-instrumental reason that does not obtain is just as wrong as knowingly harming for no reason at all. We're not permitted to gamble when so much of moral worth is at stake.

Granted, there are rare cases in which a censure poses no risk of overall harm. We might know in advance that a censure will likely harm an agent *ceteris paribus* but not *mutatis mutandis.* Consider Martin Shkreli, the 'pharmabro' who raised the price of life-saving medication solely for profit. It seems that the harms inflicted on Shkreli by the vehement censures he receives are counterbalanced by his enjoyment of the attention. Suppose that's right. (i) applies even in such cases. Two interrelated considerations explain why.

First, we routinely say that those who unjustifiably consider some individual bad, evil, or worse - deserving of censure - are being *unfair* to that individual. We consider such attributions unfair even when they are whispered outside of the earshot of their target, or even after the target is dead.[70] These epistemically unjustified attributions of negative moral worth are unfair to agents in the same way that epistemically unjustified attributions of ugliness are unfair to works of art. The person who says "though I've only glanced at it, 'The Lovers' is ugly" is being unfair to that work of art. They are not harming it. Nor are they harming Rene Magritte. Simply, fair attributions of value require the benefit of the doubt.

---

[70] In conversation, Nick Sturgeon points out that this may be unfair simply because it objectively harms the individual. If so, this is another way in which censuring on the basis of an epistemically unjustified blameworthiness judgment risks overall harm impermissibly. However, I take it that the explanation I offer above shows an additional way in which these attributions are unfair.

Second, as Peter Strawson points out, each moral agent has a level of concern for moral considerations. This level of concern constitutes each agent's 'quality of will'.[71] Individuals who make blameworthiness attributions without adequate epistemic justification don't care enough about whether or not a moral reason to censure obtains; they're willing to believe it does in the absence of confirming evidence, perhaps even in the teeth of contrary evidence. To *both* manifest this particular deficiency in one's quality of will and to deny an agent the aforementioned benefit of the doubt forbids one to censure on the basis of blameworthiness. That is so even when that censure causes no overall harm, and even if, unbeknownst to us, the agent actually *is* blameworthy. The impermissibility of epistemically negligent censures cannot be trumped by considerations unknown.

We need, then, to take stock of our degree of epistemic justification for believing agents blameworthy. Recall that, when a wrongdoer is *blameworthy*, the wrongdoer's responsibility for wrongdoing (her freely, knowingly, and inexcusably doing wrong) all by itself offers a defeasible non-instrumental justification to censure them. Debates about blameworthiness nearly always focus on whether or not agents can be responsible, with a sizable minority arguing we should be skeptical that agents have the requisite free will[72] and epistemic sensitivity.[73] As mentioned at the outset, I'll assume that the non-

---

[71] Strawson 1982. Also see Arpaly 2006; McKenna 2012, 59.

[72] Honderich 1988; G. Strawson 1994; 2011; Pereboom 2014.

[73] Rosen 2004; Levy 2009.

skeptics are correct that we *know* some wrongdoers have the freedom[74] and epistemic sensitivity[75] requisite for their responsibility. I wish to focus instead on sources of epistemic justification for credence in the proposition that

> *B*: a wrongdoer's responsibility alone offers a defeasible non-instrumental justification to censure them.

Over the next two sections, I argue that no one has adequate epistemic justification for belief in *B* (or any of its analogues). There are a few potential sources of epistemic justification for *B*. One is intuition. The other is a plausible moral theory that supports it. The former source of justification is inadequate. The latter does not exist. The absence of these adequate justifications undermines any other potential sources of justification, for example reliable testimony or peer agreement. I argue for each conclusion in turn.

2. The inadequacy of intuition.

For a proposition *P* to be 'intuitive' is for *P* to seem intellectually credible independently of an introspectively discernible inference.[76] Many, myself included, find *B* intuitive. Consider the following case.

---

[74] As Seth Shabo (2012b) points out, although all compatibilist defenders of free will argue that some agents have the free will required for blameworthiness, some of them clearly do *not* mean blameworthiness in the basic desert sense, while others are simply unclear. For compatibilists who clearly mean to defend the free will required for blameworthiness in the basic desert sense, see Fischer 2006, 6; McKenna 2012, 164-172. For libertarian defenses see Ginet 1990; Kane 1996; O'Connor 2000.

[75] FitzPatrick 2008.

[76] Audi 2008, 476-477; Sinnott-Armstrong 2006, 342.

> *Gene*: After the nation briefly falls into anarchy, the beloved revolutionary General Gene brutally tortures and executes an innocent young boy solely for his own sadistic pleasure. Gene knows it's inexcusably wrong. But he freely does it anyway.

We are epistemically justified in morally disapproving of Gene. And we are naturally angry at him in virtue of this moral disapproval. In a word, we blame Gene for what he has done. That is not enough, though, to give Gene what he deserves. In response to these gruesome wrongdoings, many of us intuit that Gene's responsibility alone offers a defeasible non-instrumental justification to express our blame in a way that harms him, that is, to censure him.[77] To isolate the intuition, assume that there's no consequentialist reason to censure Gene. He's too cautious to ever commit such risky wrongdoings again. And because only a few people will ever believe he is responsible for the wrongdoing, censuring him will be generally perceived as unjustified, deterring no one from crime, depriving the nation of one of its emerging political saviors, and thereby resulting in more bad than good.

Also, Gene's government is dissolved. There are no laws. There are no penal institutions. He is a revolutionary. He has agreed to nothing. Perhaps Gene *would* agree to such laws if he were in conditions conducive to rational decision-making, such as Rawls's original position.[78] Or perhaps, as Scanlon contends, those motivated to find

---

[77] Feinberg 1970, 55-94; Gibbard 1990, 42; Wallace 1994, 76-77; Fischer and Ravizza 1998, 5-8; Strawson, G. 2002, 452; Bennett 2002; Clarke 2005, 21; Fischer 2006, 63; Darwall 2006 and 2009; Shoemaker 2007 and 2011; Waller 2011, 2-5; McKenna 2012, 150; Boxer 2013; Pereboom 2014, 2; Rosen 2015.

[78] Rawls 1991, 102-171; 2001, 80-134.

principles for the general regulation of the behavior of others, similarly motivated, could not reasonably object to principles that allow such treatment.[79] But as Scanlon has recently conceded, our intuition is entirely independent of these contractualist proposals.[80] Moreover, as Derk Pereboom points out (in conversation), those proposals fall prey to the following dilemma. An agent in ideal contractualist conditions will agree to censure wrongdoers on the basis of *B* either because s/he considers doing so to yield the best consequences, or because of the intuition that *B*. If it's the former, then the justification for believing *B* is pragmatic rather than epistemic (an issue I address in section 4). If it's the latter, then contractualist proposals beg the question. Still, it seems to many that Gene's responsibility alone offers a defeasible non-instrumental justification to censure him, despite the absence of consequentialist or contractualist reasons to do so. He deserves a censure *just* because he freely, knowingly, and inexcusably did wrong.

Very few contemporary philosophers explicitly claim that this intuition all by itself provides adequate epistemic justification for *B*.[81] However, many seem to make that assumption. Most of those in the free will debate, for example, clearly think that the results of the debate will settle the question of whether or not agents are actually ever blameworthy. But their evidence for *B* often consists solely of thought experiments that elicit intuitions in its favor. Even some of those who express doubt about *B* are reluctant to deny the justificatory force of intuitions for it. For example, in an otherwise exhaustive series of criticisms of blameworthiness, David Boonin merely notes his own doubt about

---

[79] Scanlon 1998.

[80] Scanlon 2013.

[81] Moore (1987, 1993), and Audi (2004, 190-191) are the exceptions.

the justificatory force of intuition before conceding that an intuition that *B* might be strong enough evidence for believing that proposition.[82]

There are those who attempt to push back against *B* by appealing to intuitions contrary to it.[83] But that fire-with-fire method operates under the assumption that the *B* intuition has epistemic force in the first place. I think the assumption is unwarranted. There is no fire. There is only smoke. While intuitions are useful for distinguishing and clarifying our concepts[84], they don't always provide evidence that those concepts actually apply. Consider the application of a contentious clinical concept. A young woman anxiously complains to her doctor of chronic numbness in her hands. On first examination, the doctor finds no physical condition that can explain the numbness. He intuits that the problem is purely psychological. That intuition provides excellent evidence that the doctor has and is applying the concept of a 'somatoform disorder', the contemporary title for 'hysteria'. His further intuitions about the more specific psychological determinants of the condition, its relation to gender, its necessary and sufficient conditions, and so on, help to clarify the concept that he's employing. That hysteria intuitions are (unfortunately) prevalent in our culture is strong evidence of a shared concept.

However, neither the intuition that a person is hysterical, nor the more general intuition that there is such a malady, provides any evidence for the corresponding beliefs. The epistemic inadequacy of hysteria intuitions is not due to the non-existence

---

[82] Boonin 2008, 92-93.

[83] Scanlon 1998, 274; 2013, 104; Hanna 2012.

[84] Both traditionalists (Goldman 2007) and naturalists (Kornblith 2015, 154) about philosophical theorizing agree on this point.

of hysteria, so conceived. There really are a loosely associated cluster of mysterious symptoms that currently have no good medical explanation, and are reported to afflict females at a much higher rate than males. Maybe those symptoms are due to hysteria, so conceived. Probably, they aren't. Regardless, we still have sufficient reason to doubt that hysteria intuitions themselves are reliable evidence of hysteria. No matter how resilient or widely shared, and no matter how consciously well-intentioned and seemingly expert the diagnosis, any given hysteria intuition is likely to be the result of an unconscious gender bias that is insensitive to the relevant clinical facts. We just can't trust the intuition alone without first eliminating the effects of the bias.

Similarly, many of us intuit that a wrongdoer's responsibility alone offers a defeasible non-instrumental justification to censure them. That intuition provides excellent evidence that our core conception of blameworthiness is the 'basic desert' sense defined at the outset. However, as I argue in the next section, it's likely such an intuition is the result of an unconscious retributive bias I call 'anger affirmation' that is insensitive to the relevant moral facts. We just can't trust the intuition alone without first eliminating the effects of this bias, a task that I argue is quixotic.

2.1 The Genealogy of Blameworthiness

As a first step in identifying the likelihood of a bias, consider the nature of our retributive emotions. When we perceive agents to be responsible for wrongdoing, we quite naturally tend to feel what P. F. Strawson called 'reactive' forms of anger towards them; resentment, indignation, contempt, and the like. Empirical work confirms this. In a review of the psychological literature on anger, Shaun Nichols and Jesse Prinz point out

that various forms of immoral actions and character traits - particularly those that are perceived as "unjust" - elicit self-reports of anger, as well as all of the features of anger, including an increase in heart rate and skin temperature, and distinctive facial expressions such as a "furrowed brow, thin lips, raised eyelids, [and] square mouth."[85] Most importantly, our perception that someone is responsible for wrongdoing elicits a *sine qua non* of anger; the retributive desire to harm the wrongdoer.[86]

We already know that there is a striking correlation between this natural retributive desire and our intuitions about blameworthiness. At least *overt* expressions of reactive anger nearly always come along with the intuition that the target of anger is blameworthy.[87] That correlation has led many to suspect that our anger is the source of our blameworthiness intuitions.[88] It is difficult to substantiate such suspicions from the armchair. As Alvin Goldman rightly remarks, "a phenomenological feature [all intuitions] share is the feeling that they come from 'I know not where'. Their origins are introspectively opaque."[89] Fortunately, there has been over half a century of research on the phenomenon of 'self-affirmation' that lends credence to the view that our retributive desire is the source of our intuition that responsible wrongdoers are blameworthy.

---

[85] Nichols and Prinz 2010, 124.

[86] Vidmar 2001, 43; Darley and Pittman 2003, 333-334.

[87] Honderich 1988, pp. 583-585; Feinberg 1970, 70-1; Scheffler 2003, pp. 69-92; Strawson, G. 2008, 90; Pereboom 2007, 119 and 2014, 128-129 claim that reactive anger necessarily comes along with the *belief* that the target is blameworthy. Michael McKenna (2012, 66) argues, as do I, that there is no necessary connection between reactive anger and the belief that the target is blameworthy. My claim here is merely that reactive anger and the *intuition* that the target is blameworthy nearly always accompany one another. Granted, we tend to believe what we intuit. But that needn't be so.

[88] Nietzsche 1998; Murphy 2007, 17; Waller 2015, 39-52.

[89] Goldman 2007, 11.

The research begins with studies of the causal impact of certain sorts of decisions on attitudes. Take a study by James Fendrich, in which test subjects were asked about their attitudes towards African Americans, and also asked to decide whether they would agree to interact with African Americans in various scenarios (e.g. during lunch, as roommates, at an NAACP meeting, etc).[90] When subjects were asked the attitude questions prior to the decision questions, the kinds of attitudes that the subjects reported weren't a good predictor of the kinds of decisions they made. There was consistency between attitudes and decisions only 37% of the time. When the subjects were asked the decision questions prior to the attitude questions, the kinds of decisions the subjects made were an excellent predictor of the attitudes they adopted. The consistency between attitudes and decisions shot up to 66%. The best explanation is that the subjects' decisions to act biased them in favor of attitudes that were consistent with those decisions.

The phenomenon of subjects being biased towards certain attitudes because of their decisions, rather than the other way around, is well-confirmed. The phenomenon is most pronounced when a few further conditions are met. It turns out that the impact of decisions on attitudes rises as the decision seems to the agent to be both *voluntary* and *difficult to retract*.[91] Indeed, most researchers ensure that the subject voluntarily and irrevocably decides to ϕ by simply inviting them to ϕ. Studies show that, once the subjects voluntarily ϕ, the impact of their decision on their attitudes is very high, often above 90%.

---

[90] Fendrich 1967.

[91] Cooper 2007, 63.

What motivates the attitude change? The last few decades of studies on the phenomenon have shown that the impact of decisions on attitudes occurs because of the agent's (apparently) voluntary decision, or consequences of the decision,[92] being in tension with that agent's overall assessment of themselves.[93] Claude Steele, the pioneer of this 'self-affirmation' theory, explains that the phenomenon always occurs in studies in which subjects commit "such self-contradictory actions as writing public essays against their beliefs, expending effort on meaningless tasks, and delivering embarrassing speeches in front of prestigious audiences."[94] These actions are "self-contradictory" in that they indicate "that one is not adaptively or morally adequate...and, as a consequence [this] motivates one to reaffirm one's adequacy."[95] Steele found in his studies that

> Lacking any better means of [affirming a positive self-assessment], subjects typically attempt to justify [their actions] by changing their beliefs or attitudes to be more consistent with their actions. For example, they state that their beliefs were not really so different from the essay they wrote or that the meaningless task they worked so hard at was not really so meaningless.[96]

---

[92] Scher and Cooper 1989.

[93] Steele and Liu 1983; Steele 1988. For an overview see Cooper 2007, 90-116.

[94] Steele 1988, 269.

[95] Steele 1988, 278.

[96] Steele 1988, 269.

As Joshua Knobe and Brian Leiter report in their more recent review of the cognitive dissonance and self-affirmation literature, Steele's finding continues to be confirmed. In psychology, "the dominant view seems to be that people are motivated to believe that their behaviors are justified and that they therefore tend to adopt attitudes that justify [those] behaviors".[97]

It turns out, then, that we are biased, subconsciously, to assume that any voluntary decision that potentially threatens our positive self-assessment is justified. Return now to our retributive nature. Because our natural retributive desire is typically both easily fulfilled and stronger than conflicting desires, it provides a strong motivation to harm responsible wrongdoers that quite often results in retributive decisions and actions. We blame.[98] We punish. We lash out. We begin doing this at a very early age, and continue to decide on the basis of our anger that we'll censure when the right circumstances present themselves. This prevalent motivation, which has throughout our lives routinely resulted in retributive decisions and behaviors, is one that we feel we make voluntarily. It's also one that threatens a positive self-assessment, as acting on it is in tension with the fundamental and well-entrenched moral belief that it is wrong to harm without justification.

As the aforementioned studies show, we are unconsciously self-affirming in that any inconsistency between our voluntary decisions and our positive self-assessment prejudices us in favor of the conclusion that those decisions are justified. When met with our strong retributive motivations and the countless retributive decisions we have made,

---

[97] Knobe and Leiter 2007, 102-103.

[98] Quigley and Tedeschi 1996; Vidmar 2001, 43.

it's highly likely that this unconscious self-affirmative process prejudices us in favor of the conclusion that there is a justification for our retributive decisions and behaviors. As detailed in the Gene case, the 'mere fact' that an agent is responsible for wrongdoing is the only putative justification that can universally apply to decisions to harm responsible wrongdoers. Other moral and even prudential justifications are often lacking. So if there's a justification for censure, it would have to be 'just because' the wrongdoer is responsible for wrongdoing. It's likely this is why our conception of blameworthiness takes the basic desert form, and why *B* in particular seems true to us rather than some other kind of retributive justification. I call this effect of anger on the self-affirmation bias 'anger-affirmation'.

2.2 Clarifying the anger-affirmation account

Before discussing ways to test the hypothesis, a few clarifications are in order. First, cautious readers will rightly note that some blameworthiness intuitions occur in the absence of anger, or indeed any emotion at all. Such cases weaken the probability that occurrent anger-affirmation is the culprit of the intuition. However, the anger-affirmative process is likely shaping our concept of blameworthiness early on in our development, heavily reinforced by agreement with peers and apparent moral authorities. So, even if occurrent anger-affirmation can be ruled out as the source of a given blameworthiness intuition, it's predictable that those who have already developed and regularly deployed a basic-desert conception of blameworthiness will sometimes dispassionately intuit that a wrongdoer is blameworthy.[99] By adulthood, the assumption has been inculcated by

[99] See Jonathan Haidt's discussion of Domasio's somatic marker hypothesis (Haidt 2001, 825).

anger-affirmation. Hence the dispassionate intuition. Tellingly, it is exactly in these calm moments that our retributive intuitions weaken enough for us to recognize the suspiciousness of propositions like *B*. The problem of retribution arises when we calmly compare prohibitions against harm to our angry calls for punishment.[100]

Second, the account does not predict that everyone's blameworthiness intuitions will take the precise form that I defined at the outset. For example, most of us these days are only defeasibly motivated by our anger. So anger-affirmation makes it intuitive that the blameworthiness justification is defeasible, as defined above. But some tend instead to be sufficiently motivated by their anger. For them, anger-affirmation makes it intuitive that the blameworthiness justification is always sufficient.[101] For example, Jeffrie Murphy argues that blameworthiness *obligates* us to censure. He also admits that he finds this intuitive because "of my own resentful and vindictive Irish nature...and my not always loving personality."[102] Conversely, some will be largely unmotivated by their anger, or will simply be less angry individuals. For them, anger-affirmation makes it intuitive that blameworthiness only makes censuring 'less bad'.[103] Similarly, for most of us, the *kind* of harm our anger always motivates us to seek is the wrongdoer's guilt. I've argued elsewhere that the predominant conception of blameworthiness conforms to that motivation. But there are certainly some whose anger motivates sadistic harms, or instead merely motivates expressions of exasperation. So some intuit that

---

[100] Thanks to Michael McKenna for pressing me on this point.

[101] See, for example, Mundle (1954, 217) and Moore (1993, 15). Some go so far as to say that the justification comes in the form of an obligation (Murphy 1971 and 1973; McLeod 1999, 193).

[102] Murphy 2007, 17.

[103] Feldman 1995; Hanna 2012; Scanlon 2013.

blameworthiness is a justification to inflict more severe forms of suffering, and some intuit that blameworthiness is merely a justification to 'vent' their anger.[104]

Lastly, my account only concerns intuitions, rather than beliefs or other kinds of attitudes. The studies cited above do not make this distinction. They test for 'attitudes', broadly defined. Intuitions and beliefs strongly correlate. But they can come apart. Take each of the propositions of the well-known non-identity problem. A choice is wrong only if it makes someone worse-off. Some choices that eventually result in future individuals living in severely restricted circumstances (e.g. the choice to deplete resources) make no one worse off. Nevertheless, such choices are wrong. One of these propositions must be false. But each proposition is likely to remain intuitive, even if we reject it. The phenomenon is typical in normative ethics. Sometimes moral propositions remain intuitive, despite their proven falsity.[105]

If the test subjects in the aforementioned studies were asked to distinguish what 'seems' true to them from what they infer on the basis of reasoning, the correlation between decisions and intuitions would almost certainly be the same. It would probably increase. However, it's unclear that test subjects would also, after reporting the intuition, be quite as willing on reflection to *believe* what they intuit. I mention this in order to make clear that, while I take our blameworthiness intuitions at this stage of our moral development to be the inevitable result of an uncontrollable and unconscious anger-affirmative process, I take beliefs to be consciously revisable. Even if we abandon belief in blameworthiness, as I will be advocating we do, it'll probably continue to seem to this

---

[104] See the example of the Pirahã below.

[105] Kagan 1989, 15.

generation that wrongdoers are blameworthy. Beliefs are revisable. Intuitions just happen.

2.3 Testing the anger-affirmation account

The anger-affirmative account predicts that those who hardly ever get angry will be far less likely to find *B* intuitive. This is a difficult prediction to test, considering that humans are wired for anger. However, there is at least one source of evidence. Daniel Everett spent nearly 20 years living with the Pirahã, a small and isolated amazonian tribe. Because living conditions were so harsh, and the members of the small tribe so interdependent, it was of paramount importance for members to keep each other healthy and flourishing. As a result, Everett writes "anger is the cardinal sin among the Pirahãs."[106] My account predicts that one of the effects of a non-angry disposition will be a lack of blameworthiness intuitions among the Pirahã. It seems that's what we find. Everett relays the following remarkable story.

> One day I decided to ask one of my main language teachers, Kaaboogí, if he would work with me. I walked to his house. Coming up the path, I noticed that Kaaboogí's brother Kaapási had been drinking cachaca. I heard Kaapási yell for Kaaboogí's little white dog to stop barking. A few steps later, only fifty feet from Kaapási's hut, I saw him raise his shotgun and shoot his brother's dog in the stomach. The dog yelped and jumped, bleeding profusely, its intestines hanging from the hole torn in its abdomen. It fell to the ground twitching and whimpering.

[106] Everett 2008, 104.

> Kaaboogí ran to it and picked it up. His eyes watered as the dog died in his arms. I feared that he would shoot one of Kaapási's dogs or attack Kaapási himself. The village stared at Kaapási and Kaaboogí-quiet except for the yelping of dogs. Kaaboogí just sat holding his dog, tears in his eyes.
>
> "Are you going to do anything to Kaapási?" I asked.
>
> "What do you mean?" said Kaaboogí, puzzled.
>
> "I mean, what are you going to do to him for shooting your dog?"
>
> "I will do nothing. I won't hurt my brother. He acted like a child. He did a bad thing. But he is drunk and his head is not working well. He should not have hurt my dog. It is like my child."[107]

Two features of the case are striking. The first is that, undoubtedly due to his upbringing in the Pirahã culture, Kaaboogí is not angry at his brother for this egregious wrongdoing. The second is that the idea that there is some reason to censure his brother is completely foreign to Kaaboogí. My claim is that the former fact explains the latter. When a person is consistently devoid of any anger to self-affirm, s/he won't find it intuitive that there is a reason to censure responsible wrongdoers 'just because' they are responsible for wrongdoing. Everett points out that the Pirahã occasionally get angry with one another, but almost exclusively in situations of infidelity between couples. Unsurprisingly, these are the only situations in which the Pirahã allow a kind of punishment.[108] Again, this is exactly what the anger-affirmative account predicts.

---

[107] Everett 2008, 101.

[108] The cheating partner must stay home all day and penitently allow the cuckolded partner to playfully whack him in the head. Everett describes the process as "involv[ing] no shouting or overt anger. The giggling, smirking, and laughter are all necessary components of the process" (2008, 103-104).

This is limited evidence. But a similar phenomenon occurs even amongst those of us who regularly feel angry and have strong blameworthiness intuitions. The anger-affirmation account predicts that our blameworthiness intuitions will be difficult to sustain when the same self-affirmative bias that makes *B* intuitive to us is met with emotional motivations that run counter to a basic desert conception of blameworthiness. That's exactly what we find. Consider the much discussed case of Robert Harris.

In 1978, twenty-five year-old Harris decided, rather capriciously, to murder two teenage boys. Mere minutes after murdering the two teens, Harris nonchalantly ate one of their hamburgers, and seemed to be "in an almost lighthearted mood. He smiled and told [his brother] Daniel that it would be amusing if the two of them were to pose as police officers and inform the parents that their sons were killed".[109] Most of us rather automatically become angry at Harris for what he has done, and intuit that he is blameworthy for it. Now consider Harris's circumstances.

> He was the most beautiful of all my mother's children; he was an angel", [his sister] said. "He would just break your heart. He wanted love so bad he would beg for any kind of physical contact. He'd come up to my mother and just try to rub his little hands on her leg or her arm. He just never got touched at all. She'd just push him away or kick him. One time she bloodied his nose when he was trying to get close to her"...Robert Harris's father was an alcoholic who was twice convicted of sexually molesting his daughters. He frequently beat his children ... and often caused serious injury. Their mother also became an alcoholic and was

[109] Watson 1987, 269.

> arrested several times, once for bank robbery...Harris had a learning disability and a speech problem, but there was no money for therapy...Harris was raped several times, his sister said, and he slashed his wrists twice in suicide attempts...Everyone in the family knew that he needed psychiatric help.[110]

I contend, along with Gary Watson, Derk Pereboom,[111] Shaun Nichols,[112] and others,[113] that the force of our anger at Harris diminishes substantially after considering these further details, and that our intuition that Harris is blameworthy substantially weakens, if not disappears altogether.

What explains the substantial weakening of the intuition? Shaun Nichols argues, convincingly to my lights, that Harris's horrible circumstances evoke sympathy, an emotion that counteracts anger.[114] That diminishes our motivation to harm Harris, which in turn removes any need to self-affirm in favor of blameworthiness. At this point, we might still be able to retain the intuition that Harris is blameworthy considering that anger-affirmation has inculcated an assumption of blameworthiness. But this is less likely because our sympathy now provides some motivation, however small, to *help* Harris against those who would attempt to harm him. To fail to help someone who apparently needs it is a potential threat to a positive self-assessment. So my account predicts that the blameworthiness intuition will be difficult to retain. Instead, if our

---

[110] Watson 1987, 272-274.

[111] Pereboom 2007, 202.

[112] Nichols 2007, 411.

[113] Kane 1996, 84; Honderich 1988, 434-435.

[114] Nichols 2007, 413-6. Also see Arpaly 2006, 31.

sympathy is strong enough, it will seem that there's a justification to *defend* Harris. Jonathan Haidt cites studies in which sympathy indeed has this kind of effect on moral intuitions.[115]

Derk Pereboom offers an alternative explanation of our intuitions in the Harris case. He argues that our blameworthiness intuitions diminish in the Harris case because, once the causal antecedents of Harris's character are made clear, we come to doubt that Harris has the sort of free will requisite for blameworthiness.[116] He argues that our anger diminishes as a result of abandoning the blameworthiness intuition, rather than the other way around.

I'm unconvinced. First, most people tend to believe that an agent's upbringing and social circumstances incline but do not necessitate an agent's immoral actions, and that agents like Harris therefore have enough free will to be responsible for their actions. Nothing in the preceding description of Harris's circumstances precludes that possibility. So although free will skeptics such as Pereboom and myself are quick to see Harris's circumstances as depriving him of free will, it's not at all clear that most people are coming to that conclusion.

Second, it is likely that neither blameworthiness intuitions nor anger would substantially diminish if the causal antecedents of Harris's immoral actions were explained purely in terms that do not evoke sympathetic emotions, e.g., "Harris's acetylcholine receptors were impaired to do an increase of activity in the amygdala, which causally ensured his immoral actions." Indeed, such 'low affect' descriptions of

---

[115] Haidt 2001, 819.

[116] Pereboom 2007, 202. Also see Kane 1996, 84.

causal antecedents have been found *not* to diminish blameworthiness intuitions about 'high affect' cases like Harris's. Shaun Nichols and Joshua Knobe gave their test subjects a low affect description of a causally deterministic world. When test subjects were asked if people could be blameworthy in such a world, they answered in the negative. Tellingly, that did not weaken test subjects' blameworthiness attributions for a 'high affect' case in which a man was described as burning his wife and children alive for no reason in that same causally deterministic world.[117]

2.4 Objections to the anger-affirmation account

Intuitionists will object more generally to my account. According to intuitionists, it is an adequate understanding of the concepts involved in *B* that generates a direct 'grasp' or 'apprehension' of its truth by way of intuition. Ross, for example, claims that a general blameworthiness principle (e.g. *any* responsible wrongdoer is blameworthy) is epistemically primitive in this way.[118] Robert Audi concurs.[119] Michael S. Moore claims that it is particular blameworthiness judgments like the one we have to the case of Gene (e.g *this* responsible wrongdoer is blameworthy) that are epistemically primitive, best explained by the sort of general principle of blameworthiness to which Ross and Audi appeal.[120]

It is highly contentious that any moral proposition is self-evident in this way. But I needn't reject the self-evidence of moral propositions wholesale. We understand the

---

[117] Nichols and Knobe 2007.

[118] Ross 1930.

[119] Audi 2004, 190-191.

[120] Moore 1987, 1993.

concepts involved in *B* at least as well when we're feeling sympathetic as we do when we feel angry. But as the Harris case indicates, *B* is not intuitive when we feel sympathetic. We 'grasp' the truth of *B* in the way that intuitionists allege only if our intuitions aren't so easily and drastically influenced by affect. For example, our intuition that 'it's wrong to harm for no good reason' - a paradigm of an allegedly self-evident moral proposition - remains strong regardless of our affect. So it's not likely that our intuition that *B* is generated by our direct 'grasp' of its truth. Anger-affirmation is a better explanation.

2.5 The insensitivity of anger-affirmation

It's likely that anger-affirmation is the source of blameworthiness intuitions. Unfortunately, no part of the anger-affirmative process is sensitive to the relevant moral facts. Our self-affirmative nature biases us in favor of the conclusion that there is a justification for our decisions without regard to evidence. So if self-affirmation is the source of our conception of blameworthiness, as is likely, it is only because that conception provides us an *ad-hoc* justification for our retributive decisions. Moreover, the anger that triggers the self-affirmative process is itself neither elicited by the putative justification-conferring force of blameworthiness, nor aids us in being sensitive to such force.

That our anger is unlikely to be elicited by blameworthiness is due to its primitive emotional underpinnings. Anger is an emotion that is just as likely triggered by a host of morally irrelevant considerations as it is by responsible wrongdoing. Road rage is an obvious example. So is anger at professional athletes for their failure to win. Consider

the variety of actions and character traits that test subjects in an early and fascinatingly thorough experiment reported elicit their anger:

- "finding books out of place"
- "unpleasant manners"
- "to have [the toe] stepped on"
- "narrow mindedness"
- "I jump at conclusions and hence am often angry without cause"
- "girls talking out loud and distracting me in study hours"
- "my health being below par"
- "being kept waiting, being hurried...density in others"
- "If I am hurrying in the street and others saunter, so that I cannot get by ...or when given a seat in church behind a large pillar"
- "A discordant note in music, especially if repeated"
- "Frivolity in others"
- "An over tidy relative always slicking up my things"
- "slovenly work, want of system, method and organization"
- "late risers in my own house, stupidity"[121]

These reports confirm what those of us who are being honest with ourselves already know from experience; morally irrelevant considerations just as often elicit our anger as apparently responsible wrongdoing. That being the case, the chances are high when

---

[121] Hall 1899, 538-539. Also see Nichols and Prinz 2010, 128.

we're angry that our anger is not elicited by blameworthiness. The only way to decrease that chance is to check whether an instance of anger is indeed elicited by blameworthiness. That is precisely how we attempt (often unsuccessfully) to restrict our anger to only target responsible wrongdoers. But even when we can confirm that our anger is targeting an agent who is actually responsible for wrongdoing, given that our anger is so often elicited by morally irrelevant considerations, the chances are high that the allegedly justification-conferring force of responsible wrongdoing cited in *B* is not among the elicitors of our anger.

Some may concede that anger is not itself elicited by any blameworthiness justification, but instead can increase our awareness of it. Unfortunately, our retributive nature generally decreases our sensitivity to relevant considerations. In a recent study of the way in which unsatisfied anger makes us 'intuitive prosecutors', Julie Goldberg *et al* first review the literature on the cognitive effects of anger. According to their review, the research has confirmed that

> there are numerous ways in which anger, once activated, degrades subsequent reasoning processes. Even when the object of subsequent judgments bears no relation to the source of one's anger, anger increases: (1) a desire to blame individuals, (2) tendencies to overlook mitigating details before attributing blame, (3) tendencies to perceive ambiguous behavior as hostile, (4) tendencies to discount the role of uncontrollable factors when attributing causality and (5) punitiveness in response to witnessing mistakes made by others.[122]

---

[122] Goldberg, Lerner, and Tetlock 1999, 781-782.

Goldberg *et al* go on to find in their own study that, when the retributive desire to harm is not satisfied, anger "activate[s] an indiscriminate tendency to punish others in unrelated situations without regard for whether their actions were intentional or not."[123] This is indication that anger decreases our sensitivity to morally relevant considerations, and increases the likelihood of making false moral judgments, including false attributions of responsibility. So it's highly likely that any given instance of anger is not increasing sensitivity to any blameworthiness justification.

2.6 Summary

To summarize, we're unconsciously self-affirmative, and naturally retributive. It's likely these features of our psychology are the reason we find a basic-desert conception of blameworthiness intuitive. If we have good evidence that these features of our psychology are not sensitive to the relevant moral facts, then we have a reason to doubt that any given intuition in favor of *B* is reliable. It's clear that these two features of our psychology are *not* sensitive to the relevant moral facts. Our self-affirmative nature biases us to adopt attitudes *ad-hoc.* And the anger at the core of our retributive nature is neither elicited by any justification-conferring force of blameworthiness, nor increases our sensitivity to such force. So we have a reason to doubt the reliability of any given intuition in favor of *B*. We just can't trust it until we can eliminate the effects of the bias.

Unfortunately, it's unclear whether there's a way to eliminate the effects of the bias. We know from Steele's influential studies that a person is much less likely to

---

[123] Goldberg, Lerner, and Tetlock 1999, 783.

readjust their attitudes to accord with decisions that potentially threaten a positive self-assessment if they are given an alternative means of retaining a positive self-assessment. That provides us a promising way of eliminating the effects of any *occurrent* anger-affirmation. The trouble is that, as discussed above, the anger-affirmative process inculcates a basic-desert conception of blameworthiness early on in our moral development. Once that conception predominates, we won't see retributive behavior towards responsible wrongdoers as a threat to a positive self-assessment. At that point, the default assumption is that we're defeasibly justified in our retributive behavior. So although providing an alternative means of retaining a positive self-assessment eliminates the possibility of an occurrent anger-affirmative process, it doesn't eliminate the effects of anger-affirmation during one's moral development.

In order to reliably eliminate the effects of the bias, we need to 'check' if *B* is likely true at all. That requires a source of support for *B* distinct from intuition. By analogy, a doctor might be able to eliminate the effects of any occurrent misogyny on his intuitions. But it's unclear how he could eliminate the effects of unconscious misogyny on the concepts and background assumptions that he developed during his clinical training without first finding an alternative and reliable source of support for the sort of diagnosis he has given. He needs to 'check' other evidence distinct from his intuition that his patient is hysterical to see if there is such a thing as hysteria.

Our reason to doubt the intuition that *B* is, then, undefeated. It's uncontroversial that, if one has an undefeated reason to doubt an intuition in favor of a proposition, then that intuition does not all by itself adequately epistemically justify belief in that proposition. Therefore,

(ii) an intuition that *B* does not adequately epistemically justify the belief that *B*.

2.7 Objections to (ii)

Michael S. Moore argues that (ii) is false in cases in which it is oneself that one intuits is blameworthy. He argues that in such cases

> there is a much lessened danger that our intuitions about desert will be tainted by the emotions of *ressentiment*. Rather, one emotion here predominates, and that is the emotion of guilt. A virtuous person would feel great guilt at violating another's rights by killing, raping, assaulting, etc. And when that emotion of guilt produces the judgment that one deserves to suffer because one has culpably done wrong, that judgment is not suspect because of its emotional origins in the way that the corresponding third person judgment might be.[124]

Moore's claim that first-person blameworthiness intuitions are unlikely to be produced by anger is insightful. However, guilt-produced blameworthiness intuitions aren't any more epistemically reliable than the blameworthiness intuitions produced by anger. We genuinely feel guilty for wrongdoing *only* when we allow the conclusion that we've done wrong to negatively impact our self-assessment. But the research shows that, once we assess ourselves negatively, we'll be more likely to adopt just about *any* attitude that is consistent with that low self-assessment. So when we feel guilty, we'll be more likely to

---

[124] Moore 1993, 26.

consider ourselves blameworthy. But there's no reason to think that this 'guilt-affirmative' process is any more reliable than an anger-affirmative process (I return to the issue of guilt in section 5).

Some may argue that all moral intuitions fall prey to the sort of genealogical critique I offer here, in which case my position commits me to the radical conclusion that intuition can *never* provide adequate epistemic justification for a moral proposition.[125] That's not so. First, although there is good evidence that emotion is the source of most moral intuitions[126], there's no reason to think that *all* moral intuitions have emotional origins. For example, Marc Hauser, Liane Young, and Fiery Cushman found that common intuitions to trolley cases persisted even in those who were severely emotionally impaired.[127] So we ought not concede to the Nietzschean conclusion that all moral intuitions are "merely a sign language of the affects".[128]

Second, not all emotional motivations conflict with a positive self-assessment. Positive emotions don't conflict with a positive self-assessment, for example. So the mere fact that an emotion plays a role in eliciting a moral intuition doesn't mean it's triggering the *ad hoc* self-affirmative bias at the heart of my critique. Last, and following from the previous point, researchers have found that positive emotions are not unreliable in the way that anger is.[129] So the mere fact that an emotion elicits an intuition

---

[125] Sinnott-Armstrong 2006.

[126] Haidt 2001.

[127] Hauser, Young, and Cushman 2008, 137.

[128] Nietzsche 1966, 187.

[129] Isen 2008.

by way of the self-affirmative bias is not sufficient to undermine the justifying force of that intuition.

3. Theories of blameworthiness.

Adequate epistemic justification often comes by way of a possible inference. The sort of inference of concern is one in which the conclusion is *B* and the set of premises from which *B* can be inferred consists of propositions non-question-beggingly distinct from *B*. The best inferential justification for *B* would be a plausible theory that entails it, or at least makes it more likely than not. However, there is no such theory.

Recall that censures have both harmful and non-harmful features. Most theories claim that an agent's responsibility for wrongdoing provides a non-instrumental justification to target them with the *non*-harmful features of censures - moral disapproval and perhaps even private feelings of anger, resentment, or indignation. I'll grant that claim at the outset. My focus is instead on how the *harm* of censures is to be justified. There are two routes to justifying the harm of censures in a way consistent with blameworthiness. One is to establish that the harm of a censure is itself non-instrumentally justified. Another is to establish that the harm of censuring responsible wrongdoers is necessary to target them with the non-harmful features of censures. A theory has what I will call a *constitutive problem* when it cannot establish the former claim. A theory has what I will call a *contingency problem* when it cannot establish the latter claim. It is my thesis in this section that no existing theory of blameworthiness has established either claim.

3.1 Desert-based Strawsonian theories

Consider first a version of Peter Strawson's influential theory of moral responsibility. Strawson rightly points out that each moral agent has a level of concern for moral considerations. This level of concern constitutes each agent's 'quality of will'.[130] According to Strawson, agents who participate in normal adult relationships with one another are sensitive to the quality of will evinced by each other's actions and characteristics. This sensitivity consists in a proneness to target an agent with harmful reactive attitudes (and other such censures) when that agent's action evinces a quality of will that is inexcusably deficient.[131]

Strawson then offers two distinct lines of argument. First, he claims that these natural retributive tendencies are so deeply engrained that giving them up is not a viable option for us. Though our retributive impulses are sensitive to reasons to restrain ourselves in particular cases, the general practice is "part of the general framework of human life, [and] not something that can come up for review as particular cases can come up for review."[132] This is false. As Galen Strawson points out, humans can and sometimes do give up their retributive practices on the basis of general concerns about free will, moral sensitivity, the self, etc.[133] In any case, this first line of argument does not offer any support for *B*. That an agent can't avoid generally censuring responsible wrongdoers may offer an excuse for doing so, but certainly not a justification.

---

[130] Strawson 1982. Also see McKenna 2012, 59; Arpaly 2006.

[131] R. J. Wallace (1994, 76-77) argues that the sensitivity may sometimes consist only in the proneness to believe it apt to target such agents with blame. Also see McKenna 2012, 25.

[132] Strawson 2008, 28.

[133] G. Strawsin 2008.

Anticipating this response, Strawson's second line of argument is that, insofar as we have a choice, we can "rationally" assess whether or not to abandon our retributive practice "only in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment."[134] Strawson is clear that he takes our retributive practice to be essential to interpersonal relationships. Without it, we can at best relate to people as mere things to be managed or controlled. That being so, it is best to retain the practice (I examine a more nuanced version of this argument in section 4).

On this argument, an agent's responsibility alone does not justify censuring her.[135] Rather, on one reading of Strawson, censures are to be justified by their conduciveness to good interpersonal relations. That is a paradigmatically instrumental consideration distinct from blameworthiness. Strawson's proposal does not, then, support *B.* Recently, Michael McKenna, developing remarks by Jonathan Bennett, has tried to develop Strawson's second line of argumentation. He argues that blame is not merely conducive to, but also *constitutive* of, the non-instrumental goodness of good interpersonal relations. The argument begins with the claim that blaming is not merely an action, but is instead an interpersonal exchange. As Bennett puts it, blame is

> an activity that superficially appears to consist in one person doing something to another - and indeed abortive or perverted forms of blame can be like this. On a deeper understanding, though, [it is] something that can only properly be done together.[136]

---

[134] Strawson 2008, 28.

[135] Though we may have to assume it does on Strawson's theory, a point I discuss in section 4.

[136] Bennett 2002, 153.

McKenna expands on this claim, arguing that blaming is a process "that begins at one end with a wrong done, that then...answers that wrong by way of some blaming practice, and that invites" an interpersonal exchange between the blamed and the blamer.[137]

The proposal faces a constitutive problem because the claim being made is merely that the *process* of blame is constitutive of good interpersonal relations. That allows the possibility that the harm of the process is itself a mere means or unfortunate side-effect. In response to the constitutive problem, McKenna argues that there are three sources of non-instrumental goodness in the harmful features of this interpersonal process. First, a wrongdoer can be harmed by the blaming process only if s/he has a "commitment to membership within the moral community" in that s/he "cares about her moral relations with others".[138] Second, the blaming process is harmful only insomuch as the blamer herself expresses her own such commitment to a moral community by "expressing morality's counsel."[139] Lastly, the harm of the blaming process is brought about by repentance and reconciliation, which sustain the bonds of a moral community. McKenna contends that "there is value, non-instrumental value, in a wrongdoer's commitment to a moral community,...in a blamer's commitment to morality's counsel", and in the resolution and reconciliation "that sustain the bonds of moral community".[140]

---

[137] McKenna 2012, 169.

[138] McKenna 2012, 166.

[139] McKenna 2012, 168.

[140] McKenna 2012, 170.

But how, exactly, are the harmful features of these alleged non-instrumental goods themselves non-instrumentally good? McKenna answers that the harms of blame partly constitute these alleged non-instrumental goods in the way that the percussion work of Philly Joe Jones constitutes John Coltrane's 1958 recording of "Blue Train". "Blue Train" is non-instrumentally good, says McKenna. Assuming that's right, the percussion that constitutes it is also non-instrumentally good.[141]

However, it is not Jones' drumming itself that is relevantly analogous here, but any harm Jones might have received from that drumming. Suppose that Jones suffered some harm during the multiple takes of vigorous drumming. (This is a plausible assumption. Drummers quite often develop carpel-tunnel syndrome and other such maladies). In order for McKenna's analogy to hold, the fact that Jones's drumming or the album on which it appears constitutes a non-instrumental good must also make it non-instrumentally good that Jones's drumming harmed him. That is not plausible. Any harm Jones suffered from his drumming could only be a means or unfortunate side-effect.[142] The same goes for harmful responses to wrongdoing in general, censures in specific. If they're somehow necessarily connected to a non-instrumentally good process or outcome, it's only as a means or a side-effect.

Strawsonian proposals also face a contingency problem in that it is not clear why harm is necessary in every case to mete out the non-harmful features of censures. Harmful responses in general, censures in specific, quite often provoke emotions within the target that are consistently shown to encourage retaliation, cloud judgment, cause

---

[141] McKenna 2012, 170.

[142] Pereboom 2014, 138.

misattributions of blameworthiness[143], and generally close the mind to relevant considerations.[144] That's not always the best means to good interpersonal relations. It's often an impediment. Moreover, non-harmful alternatives seem to abound. Consider expressions of sadness, frustration, disappointment, or even cool detachment, each of which can non-harmfully communicate all that's good about a censure, and propel a wrongdoer towards the moral reform and reconciliation that McKenna, Bennett, and Boxer emphasize.[145] Consider also the dispassionate communication of rationally persuasive reasons for the wrongdoer to reform and reconcile. Surely these non-harmful measures are sometimes just as conducive to good interpersonal relations, if not more so.

How might a Strawsonian respond? Recall that McKenna highlights three steps essential to good interpersonal relations after a wrongdoing: the responsible agent's concern for her moral relations with others, her moral community's expression of 'morality's counsel', and repentance and reconciliation. McKenna and Bennett claim that each step is necessarily harmful. But that's not the case. The accomplishment of this last step - repentance and reconciliation - can feel good, especially when it is voluntary rather than coerced. And the first step - an agent's concern for her moral relations with others - can be accomplished by way of the wrongdoer's eagerness to rectify matters with her moral community, as well as her hope that she can become a better person. So there needn't be any harm at the first and third steps.

---

[143] Keltner, Ellsworth, and Edwards 1993; Goldberg, Lerner, and Tetlock 1999; Nichols and Prinz 2010, 126.

[144] Bodenhausen, G. Sheppard, L. Kramer, G. 1994.

[145] Pereboom 2014.

The second step is supposed to be harmful due to the expression of morality's counsel. If 'expressing morality's counsel' just *means* censuring, then the accomplishment of this step is indeed harmful. But then it is unclear, for reasons adduced above, why such a step is required for good interpersonal relations. Karin Boxer, responding to this kind of worry, argues that our justified negative attitudes towards wrongdoers rightly carry with them a commitment to retroactively opposing wrongdoing. According to Boxer, carrying out this commitment requires harming the responsible agent. That is one way that expressing morality's counsel might require harming the responsible agent.[146]

However, one needn't harm in order to oppose a wrongdoing. One can retroactively oppose an agent's wrongdoing by, among other things, non-harmfully expressing one's condemnation for the wrongdoing, attempting to non-harmfully reform the agent and other wrongdoers, expressing sorrow and remorse for the wrongdoing, attempting to repair the wrong in whatever ways possible, and by way of non-violent, constructive protest against the wrong. The Civil Rights Movement was marked by peaceful and highly effective opposition to wrongdoing. So one needn't harm in order to oppose wrongdoing.

Strawsonians have not provided an explanation as to why the goodness of interpersonal relations always offers a justification to censure responsible wrongdoers, much less why such a justification would be non-instrumental. So these theories cannot show that an agent's responsibility for wrongdoing all by itself offers a non-instrumental justification to censure that agent.

---

[146] Boxer 2013, 127-133.

3.2 Expressionist theories

'Expressionist' theories of punishment fair even worse in these regards. Such theories rely on the following argument.

1) A wrongdoer's responsibility alone offers a defeasible non-instrumental justification to express condemnation in a way that is proportional to that agent's wrongdoing.
2) The only way to do that is to censure.
3) So, a wrongdoer's responsibility alone offers a defeasible non-instrumental justification to censure them.

Why (1)? Igor Primoratz, expanding on remarks by Joel Feinberg[147], claims that it's because we are under obligations to "vindicate the laws [wrongdoers] broke, reaffirm the rights of their victims which they violated, and demonstrate...that their deeds were indeed crimes."[148] Just like the relation between blame and interpersonal relations in Strawsonian theories, condemnation seems merely instrumental to achieving these goals. This is why, as H.L.A. Hart remarks, any expressionist theory will "tremble on the margins" of consequentialism.[149] But even if (1) offers a non-instrumental reason to express the condemnatory features of censures, there is no support for (2).

---

[147] Feinberg 1965.

[148] Primoratz 1989, 203. Also see Feinberg 1965; Kleinig 1991.

[149] Hart 1968, 235.

Note that, if every responsible wrongdoer absconded to a deserted tropical locale, we would still have the ability to fully express our negative attitudes *to each other* in a way that vindicates laws, reaffirms rights, and demonstrates the wrongness of their actions. If anything, the absence of wrongdoers makes such expressions easier. So it's difficult to see how the accomplishment of the aforementioned goals requires censuring wrongdoers. Expressionists respond that the point is not merely to express our negative attitudes to each other, but also to guilty wrongdoers. However, as Joel Feinberg notes, punishment and other such censures seem to be merely one conventional way to express our negative attitudes to wrongdoers.[150] Thomas Scanlon presses the point succinctly. "Insofar as expression is our aim, we could just as well 'say it with flowers' or, perhaps more appropriately, with weeds."[151] So expressionist theories face a contingency problem. They don't show that censuring guilty agents is in any way necessary to express the non-harmful features of censures.

Primoratz responds by limiting the scope of (2) as applying only to those wrongdoers who are, "regrettably, although perhaps not surprisingly...oblivious to mere words [or weeds]. They do not care for the standards of society...They are lacking in respect for others...They are deficient in human sympathy... But they are endowed with as lively an appreciation of their own interest as is everyone else. So if society's condemnation of their misdeeds is really to reach them, if they are really to understand how wrong their actions are, it will have to be translated into the one language they are

---

[150] Feinberg 1965.

[151] Scanlon 1988, 214.

sure to understand: the language of self-interest. This translation is accomplished by punishment".[152]

As explained in section 1, blameworthiness is universal: it applies to any given responsible wrongdoer. Primoratz's limitation violates this constraint. In any case, Primoratz is highlighting not just the expression of negative attitudes to guilty wrongdoers, but its effective *communication*.[153] The communication of negative attitudes to a wrongdoer is never necessary to vindicate laws, reaffirm rights, or demonstrate wrongdoing. Moreover, those lacking in sympathy and respect are exactly the sorts who will respond most negatively to harmful treatment. Their sympathy, respect for persons, and sensitivity to the wrongness of certain actions is likely to decrease rather than increase. So Primoratz's own explanation entails that premise (2) will not apply to those specific wrongdoers. But those are the only cases that he alleges (2) applies to. So Primoratz has not shown that the expressionist theory provides any reason to prefer that a response to wrongdoing comes in the form of a censure rather than some non-harmful alternative.

Jean Hampton argues for a version of the expressionist view that focuses on the relative moral status of victimizers and victims. On her view, wronging a person always involves implicitly declaring oneself, falsely, as having a higher moral status than that person. In that case, "the retributivist demands that the false claim be corrected. The wrongdoer must be humbled to show that he isn't the lord of the victim."[154] The only way to do that, according to Hampton, is by way of punishment. "Retributive punishment is

---

[152] Primoratz 1989, 199-200.

[153] For a discussion and critique see Hanna 2008.

[154] Hampton 1991, 398.

the defeat of the wrongdoer at the hand of the victim (either directly or indirectly through an agent of the victim, e.g. the state) that symbolizes the correct relative value of wrongdoer and victim. It is a symbol that is conceptually required to reaffirm a victim's equal worth in the face of a challenge to it...[155]

Unfortunately, as David Dolinko points out, even if the falseness of a claim implicit in an instance of victimizing were itself what makes that victimization wrong, it wouldn't provide a justification to censure. "If implicit assertion of that false message were what made the act wrong, explicit assertion of the falsehood ought to be every bit as wrong (and deserve just as much punishment) as its implicit assertion through the criminal act." However, that's clearly false. "If someone publishes a book asserting that...its author is an *Uebermensch* greater in moral value than any other human being on the face of the earth, we do not regard it as obligatory...to clap the author in jail."[156] Moreover, a false claim of moral superiority is not what makes victimization wrong. A deluded cultist may believe in his religious fervor that asking his toddlers to join him in ritual suicide is a means of recognizing their high moral status. That the cultist's victimization of his children contains no false claim of moral superiority does not make that victimization any less wrong.

One response available to Hampton is that the false claim of moral superiority is not one that is always *endorsed* by victimizers, but rather one that victimization necessarily evinces. This is the reading that Heather J. Gert, Linda Radzik, and Michael Hand offer in their excellent review of Hampton's work on retributivism. They read

---

[155] Hampton 1991, 398-399.

[156] Dolinko 1991, 551.

Hampton as saying that, "When one person wrongs another...the offender does not merely express the claim that she is more valuable than the victim. She has actually dominated the victim, thereby providing substantive evidence for her claim to superiority."[157]

This line of response is not promising. First, very few victimizations indicate to anyone that the victimizer is somehow superior. We see victimizers as lowly characters, and their victimizing actions as confirming evidence of that low status. If they deserve a censure, they deserve it for victimizing, not misleading. Second, even if victimization did provide misleading evidence, the only reason to defeat that evidence would be instrumental. The victimizer would deserve a censure, not due to the wrongdoing itself, but rather as a means to counteract the misleading evidence that the wrongdoing allegedly provides. Lastly, in order to defeat misleading evidence of a victim's low moral worth, we would need to make explicit the morally salient attributes of the victim. Censuring the victimizer is not the best means to that end. Hampton herself, elaborating on remarks by Feinberg, acknowledges that in some cases a parade for the victim is better than a censure as a means of defeating the false evidence of low moral status.[158] It's rare anything so elaborate would be required. Often, just a bit of reflection provides sufficient evidence of moral worth.

3.3 Fair play theories

---

[157] Gert, Radzik, and Hand 1994, 82.

[158] Hampton 1992, 1697.

Consider next a 'Fair Play' theory of punishment, which George Sher claims can be amended to explain blameworthiness.[159] Herbert Morris, the progenitor of the view, points out that a benefit offered by any just society is the benefit of noninterference by others in life and bodily security. I'll call this the *safety benefit*. According to Morris, a burden we voluntarily accept in order to receive this benefit is "the exercise of self-restraint...over inclinations that would, if satisfied, directly interfere or create a substantial risk of interference with others in proscribed ways".[160] I'll call this the *safety burden.* It follows that any person who freely and knowingly[161] eschews this self-restraint "renounces a burden which others have voluntarily assumed and thus gains an advantage which others, who have restrained themselves, do not possess".[162] That, claims Morris, yields an unequal distribution of benefits and burdens. Criminals get safety benefits without safety burdens.

According to Morris, all else being equal, we ought to ensure an equitable distribution of safety benefits and burdens. So, all else being equal, we ought to rectify any unequal distribution caused by criminal wrongdoing. However, we ought to do so in whatever way violates the least rights. Morris argues that censuring criminals is the least violative way to return to an equitable distribution.[163] Hence, all else being equal, we ought to censure criminals. More formally,

---

[159] Sher 1989, 69-90.

[160] Morris 1968, 477.

[161] Morris makes the freedom condition clear. "A person has not derived an unfair advantage if he could not have restrained himself or if it is unreasonable to expect him to behave otherwise than he did" (478). The epistemic condition I offer him for the sake of charity.

[162] Morris 1968, 477.

[163] Morris 1968, 477-478.

1) All else being equal, we ought to ensure an equitable distribution of safety benefits and burdens in the least violative way possible.
2) Those guilty of criminal wrongdoing inequitably benefit from their wrongdoing.
3) Censuring them is the least violative way to return to an equitable distribution.
4) So, all else being equal, we ought to censure criminals.

I do not interpret Morris himself as offering a theory of blameworthiness. Nor do I think any premise of the argument can be adapted for that purpose. Consider (1). Morris seems to think that we ought to equitably distribute safety benefits and burdens because it is in our self-interest to do so. That's a highly plausible explanation. It's also a paradigm of contractualist reasoning. But if (1) is indeed a contractualist principle, then the fair play theory is not a theory of the sort of desert-based blameworthiness at issue. No alternative explanation seems forthcoming.

Also, blameworthiness applies to *all* wrongdoing. So the argument cannot explain blameworthiness unless it goes beyond mere criminal wrongdoing.[164] However, (1) is implausible if we take it to apply to all wrongdoing. We indeed ought to set up social institutions that protect us against criminal harms (murder, mugging, fraud, theft, and the like), and that actively force upon everyone the 'burden' of not carrying out those harms. When done right, police-work is essential. But neither individuals, nor social institutions, nor the moral community have a reason to provide protection from *every* kind of wrong, much less force upon every agent the 'burden' of never wronging others. Consider the

[164] Sher 1987, 78.

wrong of privately disparaging a close friend, resenting a parent, or neglecting to feel the appropriate amount of love for one's child. No one has the moral standing to preemptively protect victims against these wrongs. No one has the moral standing to stop the wrongdoer in advance. Some morals oughtn't be policed.

As for (2), how is it, exactly, that wrongdoers benefit from wrongdoing? Morris cites 'giving in' to certain inclinations or impulses that others do not as a sort of benefit. Richard Burgh, in an otherwise insightful critique of the fair play theory, claims that this is indeed a benefit in that it affords the wrongdoer "a bit more freedom than those who undertook the [safety] burden."[165] George Sher agrees.[166] It is difficult to see how this alleged freedom constitutes a benefit. Suppose that, during the course of otherwise great lives, brothers Jim and John each develop momentary inclinations to strike their frustratingly senile father. Jim tragically gives in to his inclination. John doesn't. Perhaps there is some understanding of 'freedom' in which Jim gained more freedom than John by giving in to his inclination. But given that acting on such an impulse can bring Jim only sorrow, there is no good explanation of why Jim 'benefits' from it. Indulging an impulse to do wrong sometimes affords us a benefit. It does not constitute one.

George Sher claims that a wrongdoer benefits in that he gains "freedom from the demands of the prohibition he violates."[167] What does it mean to be free of the demands of morality? Sher's remarks are brief. He claims that, "because others *take that prohibition seriously*, they lack a similar liberty" (emphasis mine).[168] This indicates that,

---

[165] Burgh 1982, 209.

[166] Sher 1987, 80.

[167] Sher 1987, 82.

[168] Sher 1987, 82.

on Sher's view, to be free of a demand of morality is to *not* take that demand seriously. In a later work, Sher specifies that an agent takes a demand to not ϕ seriously if that demand *qua* demand demotivates her against ϕ-ing. So, those who don't take a demand seriously lack at least one demotivation to violate that demand. Lacking a demotivation against an action is a kind of freedom. Moreover, since moral prohibitions come in various degrees of strength, freedom from moral demotivation also comes in degrees.[169]

There would seem to be only two ways that this can benefit an agent. It can allow the agent to acquire a benefit in some specific instance, or more easily acquire benefits in general.[170] Lacking a demotivation to thieve, for example, offers a specific benefit whenever there's an opportunity to snag a stack of cash. Those who don't worry themselves about morality in general may be benefitted in certain societies. As Sher says, "each increase in an agent's willingness to disregard the requirements of morality leaves him in a better position to accomplish the goals that partially determine his interests."[171] Unfortunately, lacking a demotivation against wrongdoing, in specific or in general, is often not beneficial. Consider those thieves who wind up with marked bills, or those imaginary societies that tend to reward the virtuous more than the vicious. When wrongdoers are in these circumstances, they don't gain the benefit under consideration. Sher's version of the fair play theory entails that those wrongdoers cannot be blameworthy. But the theory is supposed to entail that such wrongdoers are blameworthy.

---

[169] Sher 1997, 168 - 173.

[170] Dagger 1993, 480-484.

[171] Sher 1997, 174.

As for (3), the problems identifying any additional benefit wrongdoers gain from doing wrong makes it difficult to assess how censuring removes that additional benefit. However, I'll set that issue aside and focus instead on the claim that censuring is the *least violative* way of rebalancing benefits and burdens. If censuring wrongdoers somehow removes a benefit, then praising or rewarding should offer a benefit. Praising or rewarding hardly ever violates anyone's rights, or leads to suboptimal outcomes. So it's hardly ever more violative to reward those who do right than to censure those who do wrong. Assuming (1) and (2) of the fair play argument, it follows that there will be some cases in which we ought to praise or reward those who abide by moral norms, rather than censure those who do not. That's a nice conclusion for those who favor positive reinforcement. It's not much help to those seeking a theory of blameworthiness.

4. An argument for retributive caution

So far, I have argued that

> (i) An agent's alleged blameworthiness permits one to censure that agent only if one has adequate epistemic justification to believe s/he's blameworthy.
>
> (ii) An intuition that an agent is blameworthy does not adequately epistemically justify that belief.
>
> (iii) There is currently no plausible theory of blameworthiness.

We can sometimes gain epistemic justification by way of peer agreement or certain forms of testimony. However, such sources are highly dubious, to say the least, when

we have sufficient reason to conclude that our peers and testimony providers also lack a reliable intuition or argument in favor of their pronouncements. (ii) and (iii) provide that reason. So,

> (iv) If (ii) and (iii), no one has adequate epistemic justification to believe an agent is blameworthy.

It follows that

> (v) An agent's alleged blameworthiness never permits censuring that agent.

As discussed at the outset, there is a widespread practice of treating responsible wrongdoers as if blameworthiness permits censuring them. According to (v), this retributive practice harms people on the basis of an alleged moral principle that we have sufficient reason to doubt. Most moral theories forbid us from adopting harmful practices that require acting on the basis of moral principles we have sufficient reason to doubt. On these 'single-tier' theories, (v) entails that we ought to practice retributive caution by refraining from censuring wrongdoers on the basis of alleged blameworthiness.

However, 'two-tier' theories have the resources to allow, even demand, adopting practices that involve acting on moral principles we have sufficient reason to doubt. I'll discuss two notable two-tier justifications for blameworthiness - one that I think is available in Peter Strawson's work, and Manuel Vargas's *Agency Cultivation Model* -

before identifying and criticizing a premise required to defend any two-tier theory of blameworthiness.

4.1 Strawson (again)

In section 3, I discussed two lines of argument in Strawson. The first is naturalist, the second consequentialist. As I mentioned above, neither was meant to provide evidence that anyone actually is blameworthy as I have described that concept. However, Strawson's second line of argument can be (and I think is best) read as a two-tier defense of blameworthiness. He is clear that we are prone to censure agents for freely, knowingly, and inexcusably manifesting a poor quality of will, and *not* for instrumental reasons. "The efficacy of these practices", he says, "is not a sufficient basis, it is not even the right *sort* of basis, for these practices as we understand them."[172] This indicates that Strawson takes our retributive practice to be roughly as I have described it; we tend to censure responsible wrongdoers *just* because they are responsible for wrongdoing.

However, recall that, according to Strawson, the question of whether or not to abandon our retributive practice can be assessed "only in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment."[173] Puzzlingly, that's an assessment of the 'efficacy' of the practice that he clearly thinks will come out in favor of retainment. Considering Strawson claims efficacy is not the right sort of basis for the practice, he might have meant this final line of defense to be two-tier. One way to

---

[172] Strawson 2008, 21.

[173] Strawson 2008, 28.

read him is that he's conceding that we tend to censure in response to responsible wrongdoing, and not in response to instrumental considerations. But when we ask whether or not responsible wrongdoing all by itself provides a justification for censuring, he's claiming that it doesn't particularly matter. The retributive practice is essential to interpersonal relationships. So it is best to retain it even if doing so involves systematically acting as if an epistemically unsupported justification obtains.

4.2 Vargas's Agency Cultivation Model

Manuel Vargas reads Strawson in this way, and argues for a similar view. Like Strawson, Vargas points out that we tend to censure directly in response to responsible wrongdoing, and not in response to instrumental considerations. When we do so, we believe it is responsible wrongdoing itself that justifies the censure. But Vargas denies that responsible wrongdoing all by itself justifies censuring.[174] It follows that, on his view, our retributive practice is likely to involve acting as if an epistemically unsupported justification obtains.[175] However, Vargas argues that our retributive practice on the whole fosters moral agency, and we're justified in retaining it in virtue of that fact, even if it involves responding to each individual case as if there is a blameworthiness-like justification for doing so.[176] We should revise our beliefs as much as possible, of course. But ultimately all that matters, Vargas argues, "is that the system as a whole produces

---

[174] Vargas 2013, 249-261. Moreover, he's skeptical that agents are responsible in the sense likely to be presupposed by our retributive practice (2013, 21-72).

[175] Vargas 2013, 97.

[176] Vargas 2013, 172; 2015, 2666-2671.

agents that, over time and in a wide range of contexts, are suitably responsive to moral considerations."[177]

4.3 Bad effects

I offer such a quick gloss of these theories only because my concerns are rather general. Two-tier theories of blameworthiness all share the following feature. They concede that our retributive practice could be epistemically faulty in that it involves responding to an apparent justification that we have sufficient reason to doubt. They then argue for retainment in any case by claiming that our retributive practice conduces to some crucial human good (e.g. interpersonal relationships, agency cultivation) better than alternatives.[178]

One strategy against this claim involves skepticism about the crucial goods at issue. Stephen Morris, for example, argues that the goodness of moral agency cultivation has not been established.[179] I'm resistant to that argumentative strategy. Another strategy is to deny that the retributive practice best conduces to the target goods.[180] This strategy is more promising. Although there are clearly cases in which censuring is conducive to some crucial human good like interpersonal relationships or agency cultivation, it's unclear that our retributive practice is the most effective means to this end. I consider the widespread belief that it is to be wishful thinking.[181]

---

[177] Vargas 2013, 176.

[178] Nichols (2007) is also sympathetic to this sort of view.

[179] Morris 2015.

[180] Again, see Morris 2015.

[181] Thanks to Derk Pereboom for introducing me to this idea.

Unfortunately, I don't have the resources to argue that our retributive practice is not the most conducive to the aforementioned goods *in the aggregate*. No one does. It's an empirical question with a paucity of data. My strategy against the two-tier view is to argue that, even if retainment maximizes the aforementioned goods, that still doesn't outweigh the damaging effects of retainment. Consider just two of these effects.

One damaging effect of retaining our retributive practice is that it involves routinely acting as if an epistemically unsupported justification obtains.[182] This involves systematic and widespread irrationality, and may cause psychological damage to those aware of it. Vargas claims it is possible, through conceptual revision, to revise basic desert out of our first-order blameworthiness judgments so that those judgments are consistent with instrumental second-order justifications.[183] However, as much of the discussion so far has made clear, the 'basic-ness' of our blameworthiness judgments is incredibly resilient. That being the case, the skeptical challenge pushes us towards an error theory about blameworthiness, not revisionism. Furthermore, revising to believe that we are justified in censuring only by virtue of instrumental justifications would leave us feeling alienated from the first-order non-instrumental conviction on which we act.[184]

Another damaging effect of retainment is its collateral damage. On the two-tier view, censuring responsible wrongdoers is justified by the *overall* efficacy of the

---

[182] Pereboom 2013, 129-130.

[183] On his view, although the retainment of our retributive practice does require judging that 'something' about responsible wrongdoing defeasibly justifies censure, these judgments are 'cognitively thin' in that "what that something comes to is...not anything about which we have formed thoughts or strong convictions. Indeed, from the standpoint of our ordinary discourse and practical life, what is important is confidence that we are correct and justified in our first-order judgments. The justifiers and their precise nature are ordinarily of secondary and considerably less importance". Hence the ability to revise first-order judgments to fit with second-order instrumental justifications (2015, 253).

[184] Doris 2015.

practice, and not by the efficacy of censuring in each particular case.[185] So there will be many instances in which our retributive practice advises censuring people even though that censure won't foster agency, improve relationships, or produce any other moral good. On the two-tier view, we still have reason to censure these individuals. They are to be ground up in the gears of the two-tier machinery.[186]

With these damaging effects in mind, it becomes clear why the two-tier justification will dissolve if there's a less damaging alternative practice that is even *close* to being as effective. By analogy, consider the largely abandoned stop-and-frisk policy in NYC. That policy systematically treated innocents who fit certain profiles as if they were suspected criminals. Advocates of the policy often claim, falsely, that it was better at decreasing crime than alternatives. These advocates fail to consider that, even if the policy were better than alternatives at decreasing crime, that would not alone justify it. The large number of innocents swept up by stop-and-frisk were collateral damage whose demographic profiles were never sufficient to warrant suspicion of guilt in the first place. A policy that did not treat innocents in this way would be morally preferable even if it were, say, 80% as effective at crime prevention.

Similarly, our retributive practice cannot merely be just a little better than the other available methods at maintaining interpersonal relations, fostering moral agency, or achieving other such goods. For if that were so, we would be paying a premium in the harms of irrationality and collateral damage to achieve marginal increases in the target goods. A system that is beneficial overall will strike a better balance between gains in

---

[185] Vargas 2013, 172.

[186] McGeer 2015, 2639-2640.

the target values and losses of others. So, our current retributive practice is justified by its effectiveness in bringing about the aforementioned goods only if there is not a less damaging alternative practice that is almost as effective.

Our retributive practice does not meet this standard. To see why, consider a kind of objection that is widely thought to push rule consequentialism towards act consequentialism. If agency cultivation, securing good interpersonal relations, or some other such good is our aim, why not just do our best to censure responsible wrongdoers when and only when it conduces to their agency cultivation, good interpersonal relations with them, etc? As Victoria McGeer points out, two-tier theories like Vargas's 'collapse' into single-tiered theories because it is far better for us to respond directly to the justifications for censure that actually obtain, rather than taking the tortuous route through epistemically unjustified beliefs about blameworthiness.[187] After all, the direct route would presumably be roughly as effective as our current practice, but without irrationality or alienation, and with far less collateral damage.

Vargas responds that, if an agent is blameworthy when and only when blame conduces to their agency cultivation (or conduces to good interpersonal relations with them, etc), then the dead, the dying, the distant, and the blame-resistant aren't ever blameworthy. But facts about such agents' blameworthiness "present themselves as settled independently of whether the considered agents are sensitizable in that particular instance. Yet McGeer's account seems to require that we conclude that this swath of ordinary practices is fundamentally flawed."[188] That's a severely counter-

[187] McGeer 2015.

[188] Vargas 2015, 2673.

intuitive implication that requires "radical revision" to our moral thought and practice. "The two-tiered picture of blaming norms", on the other hand, requires far less revision because it "provides a way of capturing the spirit of the idea that 'praise and blame are justified not in terms of their putative effects (whether good or bad), but simply because they are the appropriate or merited response to what the person has done'."[189]

Radical revision to ordinary moral thought is indeed required to adopt a single-tiered theory. However, the fact that a moral theory conflicts with ordinary moral thought is a demerit of that theory only if there is some reason to think that ordinary moral thought is correct. A claim at the core of Vargas's view is that our ordinary moral thought about blameworthiness is *not* correct. Surely, then, the fact that responding directly to the bona fide moral reasons that Vargas identifies requires radical revision to epistemically unjustified moral beliefs does not weigh in favor of retaining those beliefs. Adopting less biased policing policies, for example, requires radical revision to many people's ordinary (authoritarian) moral thought. That alleged 'cost' of revision doesn't come close to outweighing the damages of retaining such a practice. We ought not stop-and-frisk, despite unjustified beliefs to the contrary. Why think that matters are any different in the case of retributive beliefs?

Perhaps what Vargas has in mind here is not just the counter-intuitiveness of responding directly to instrumental reasons for censure, but also the difficulty of it. He claims his two-tier agency cultivation model is "buttressed against such a collapse [into a single-tiered theory] by the realities of our psychologies, including the length of time it takes to develop and refine moral attitudes, the flexibility of our attitudes, the cognitive

---

[189] Vargas 2015, 2675.

burden involved in assessing responsibility, and the overarching need to have a stable and efficacious system of influence. Collectively, these considerations...weigh against something like act or token specific norms of responsibility."[190]

Vargas's remarks indicate that he takes it to be too psychologically difficult for human agents to efficiently and systematically censure on the basis of instrumental justifications on a case-to-case basis. This is speculative. As I argue elsewhere[191], it is psychologically possible, even easy in some kinds of cases, to censure solely on the basis of such justifications. We often censure young children, for example, when we feel it will best morally educate them. We censure immoral but nevertheless undeserving agents in order to overcome or take a stand against the immoral systems of which they are a part. Perhaps we can, with slow and steady practice over several generations, replace our retributive practices with the forward-looking approach that we apply in these sorts of cases. There is not yet sufficient reason to dismiss this possibility, especially considering the costs of retainment.

Since an agent's alleged blameworthiness does not permit censuring that agent, and there is not sufficient reason to act as if it does, we ought to practice *retributive caution* by refraining from censuring agents on the basis of blameworthiness.

5. Censures, blame, and guiltworthiness

My argument has focused only on the limited class of blaming responses I call censures. This leaves room for the following objection. At the outset, I identified blame

---

[190] Vargas 2013, 181.

[191] See "Skepticism and the Rationality of Anger".

as anger by virtue of moral disapproval. Well, we are clearly epistemically justified in morally disapproving of those who are responsible for wrongdoing. And it is at the cornerstone of my account that we are prone to become angry at, and hence blame, these individuals. On my view, blame doesn't need to be expressed in a way that is harmful, or even expressed at all. So, even if I'm right that we ought to be skeptical that *harmful* expressions of blame (i.e. censures) are deserved, my account allows that blame itself is deserved by these wrongdoers. Isn't this harmless blameworthiness a sort that is allowed by my account?

Perhaps not. Gary Watson remarks, plausibly, that "one's blaming attitudes are unfair if it would be unfair...to subject others to the adverse treatment to which one's attitudes dispose one".[192] If he's right, the impermissibility of censuring on the basis of blameworthiness entails the impermissibility of blaming on the basis of blameworthiness. However, I feel no need to push this point because I find the idea of harmless blameworthiness implausible. As I argue elsewhere,[193] debates about *what* is deserved by wrongdoers have for too long suffered from an unproductive obsession with blaming attitudes and emotions, when what is clearly at stake is the harmful effects of acting on them. Positions that prohibit desert of the latter and allow desert of the former makes this all the more clear. Such a position entails that people are justified in fuming to themselves over the transgressions of others, but not justified in doing much of anything about it. This action is, as the old saying goes, like drinking poison and expecting another to sicken. It is not our attitudes and emotions that responsible

---

[192] Watson 1996, 239.

[193] "Moral Responsibility as Guiltworthiness".

wrongdoers intuitively deserve, but rather some of the effects that typically result from acting on these attitudes and emotions.

However, consideration of this point leads us to a kind of harmful response that I think escapes my argument for retributive caution. One of the typical and intended harms of overtly blaming an individual is the eliciting of moral guilt in that individual. This emotion is not, as much of the current literature suggests, a form of blame. Blame consists of some form of anger by virtue of moral disapproval. Moral guilt (henceforth simply 'guilt') consists instead of sorrow by virtue of moral disapproval of oneself. The difference is important. Anger aims to hurt. Sorrow does not. It just hurts. Anger clouds thoughtful contemplation. Sorrow does not (as socially isolated academics are acutely aware). It is often the case that we judge that we *deserve* to feel this non-hostile emotion for our wrongdoings. I do not think that intuitions to this conclusion are much more trustworthy than blameworthiness intuitions. But the following inference is readily available to each of us and is, I think, often employed.

It's partly constitutive of being a morally good person to feel guilt for one's responsible wrongdoing. I have a non-instrumental justification to be a good person. Therefore, I have a non-instrumental justification to feel guilty for my responsible wrongdoing. The argument is typically run from this first-person point of view. But it applies generally. If the desert of guilt for responsible wrongdoing is justifiably applied to oneself, it's justifiably applied to others as well.[194] *Each* of us has a non-instrumental justification to be a good person, and hence a non-instrumental justification to feel guilty for our responsible wrongdoings.

---

[194] Moore 1993, 27.

As far as I can tell, the apparent plausibility of the argument does not result from an epistemically unreliable process. Nor does the argument fall prey to the criticisms I leveled throughout section 3. So although we are not epistemically justified in concluding that anyone is blameworthy, we may be epistemically justified in concluding responsible wrongdoers guiltworthy. This conclusion leaves room for censures to be defeasibly justified whenever they are the best means to elicit deserved guilt. That instrumental reason for censure, together with the others mentioned throughout this essay, show that retributive caution does not commit us to retributive prohibition. Retributivists might find this concession patronizing. They are welcome to censure me, provided they can identify an instrumental reason to do so.

Moral Responsibility as Guiltworthiness

A common way to explain the 'liability' sense of moral responsibility for an agent's immoral action or characteristic (henceforth 'transgression')[195] is by the kind of negative response that the agent deserves for it. The predominant view amongst those who explain moral responsibility in this way is that blame is always the kind of response at issue.[196] Stronger responses - punishment, public censure, and the like - may also be deserved, but not always, and only if blame is deserved.[197] I call this the *blame-focused view of moral responsibility*. In this paper, I argue for an alternative. Aside from a few exceptions[198], moral guilt is cited in recent discussions of moral responsibility, if at all, as merely an effect of being blamed[199], a form of blame[200], or as a reliable indicator of moral responsibility[201], but not itself the kind of response that explains moral responsibility for transgressions. Getting behind the eyes of immoral agents reveals otherwise. Moral responsibility for transgression is explained by the way in which the transgressor deserves to feel, and not by the interpersonal responses that typically elicit those feelings. Moral responsibility is intrapersonal. Guilt, not blame, is the paradigm.

---

[195] Below, I more carefully consider moral responsibility's relation to obligations.

[196] Feinberg 1970, 55-94; Gibbard 1990, 42; Wallace 1994, 76-77; Fischer and Ravizza 1998, 5-8; Strawson, G. 2002, 452; Bennett 2002; Clarke 2005, 21; Fischer 2006, 63; Darwall 2006 and 2009; Shoemaker 2007 and 2011; Waller 2011, 2-5; McKenna 2012, 150; Boxer 2013; Scanlon 2013, 101- 102; Pereboom 2014, 2.

[197] See especially Feinberg 1965 and 1970, 55-94.

[198] Gibbard 1990; Clarke 2013.

[199] Bennett 2002, 152-153; McKenna 2012, 161-171.

[200] Morris 1988; 66-67; Wallace 1994, 66-67; Skorupski 1999, 142; Darwall 2006, 112; McKenna 2012, 72-74.

[201] Moore 1993.

Blame and other such interpersonal responses are like a police composite, accurately portraying a common face[202] of moral responsibility without identifying its DNA.

My plan is as follows. In section 1, I discuss the nature of blame, moral guilt, and distinguish moral responsibility-based justifications to respond to transgressors from other justifications to regard or respond. I conclude the section by arguing that an agent deserves to feel moral guilt if she deserves blame. In section 2, I argue that the converse does not hold. Some transgressors do not deserve blame, and yet still deserve to feel guilt for their transgressions. Denying that those transgressors are morally responsible is implausible. So it is guilt rather than blame that is always the kind of response that morally responsible transgressors deserve. I call this the *guilt-focused view of moral responsibility*. I argue that this alternative view conforms to general features of desert that the blame-focused view does not, offers a compelling way to reconcile conflicting intuitions about the suberogatory, and allows that those outside of the moral community can still be morally responsible for their transgressions. I conclude in section 3 by considering objections.

1. Disapproval and blame.

Gary Watson argues that there is a sort of 'aretaic' blame that consists merely in dispassionate moral attitudes towards an agent.[203] The attitudes necessary and

---

[202] The metaphor originates with Watson's influential 1996 paper "The Two Faces of Responsibility".

[203] Watson 1996.

sufficient for this sort of blame are contested. But it is clear that it consists in some number of the following.[204]

(i): the belief that the agent is or has acted in a way that is morally bad.

(ii): the belief that negatively appraising the agent for this transgression is apt.

(iii): an endorsement of the moral reasons against the agent's transgression.

(iv): desiring that the agent not have transgressed.

(v): holding the desire cited in (iv) because of (i)-(iii).

There is a noteworthy dispute concerning the use of the term 'blame' here. Michael McKenna has us suppose that a person who holds attitudes (i) - (v) has no negative emotions or ill feelings towards their target and "could not imagine regarding or responding to [the target] any differently than she would her sweet little grandmother... [believing] that there's no better remedy than just hoping [the target] will do better next time".[205] It seems to me that such an individual is not blaming *per se*. Rather, I think that the holding of these attitudes are better described as a sort of *moral disapproval* that is a precursor to blame. That is how will refer to them for the remainder of this work. I take it that in paradigm cases to genuinely *blame* an agent for an immoral act consists in

[204] For example, Arpaly (2006, 9-39), Smith (2007, 476), and Scanlon (2008, 122-214) argue that (i) through (iii) are necessary and sufficient. George Sher argues for a view of blame that only requires (i) and (iv) (2006, 93-114). Michael McKenna argues that all of these conditions are required (2012, 21-29).

[205] McKenna 2012, 24. Also see Rosen 2015, 67.

Blame: some form of anger (such as resentment or indignation) towards an agent in virtue of morally disapproving of that agent.[206]

One might feel other negative emotions by virtue of moral disapproval. Some, such as frustration and irritation, will in a variety of contexts be relevantly blame-like, perhaps counting as a 'mild' form of blame. I am happy to also include them in the category. I take other emotions, such as disgust and hatred, to be in most contexts too unlike blame to fall in the category.

1.2 Two kinds of moral responsibility

For the purposes of this work, this particular dispute over the nature of blame is merely terminological. Although I reject that the collection of attitudes that I call moral disapproval constitutes a kind of blame, I take them to be pertinent to moral responsibility in exactly the ways that Watson and others claim. To morally disapprove of an agent is not merely to attribute to them some kind of causal responsibility, for example when we say 'James is responsible for the fire because he accidentally dug into a hidden gas line.' To morally disapprove of an agent for a transgression is to take that agent to be in a crucial sense *morally* responsible for that transgression due to immoral ends that s/he has adopted. Specifically,

---

[206] Strawson 1982; Wallace 1994, 76-77; Watson 1996, 238; Bennett 2002 and 2008; Smith 2007, 477; McKenna 2012, 21-29; Boxer 2013; Rosen 2015. Gibbard (1992) argues non-cognitive anger is sufficient for blame.

> An agent is *morally responsible in an attributability sense* for a transgression just in case it is apt to morally disapprove of that agent for the transgression.[207]

Moreover, I agree with Watson and numerous others that an agent's attributability does not entail that the targeted agent is liable to a response such as blame or punishment. Unlike blame and punishment, moral disapproval is not a way of responding to or treating agents. It rarely, if ever, affects targeted agents in any way, and certainly never harms[208] them in ways that are typically prohibited. This is why, as George Sher points out, moral disapproval is made apt merely by those norms "that require that we believe propositions that are true and that we accept moral principles that are justified."[209] For example, suppose that,

> *Janus's attributability*: After just a few drinks, your good friend Janus relays a deeply embarrassing secret of yours to some mutual friends. You're mortified. Your friends will never see you the same way. The next morning, Janus admits that he likes to gossip, and had a small urge to do so that night that is characteristic of an immoral value that he holds. But he claims that his acting on this urge was an unexpected and unintended result of the mixture of his newly

---

[207] Zimmerman 1988, 152; Watson 1996; Arpaly 2006, 9-39; Sher 2006, 130; Shoemaker 2011; Nelkin 2011, 34-35.

[208] Throughout, I assume that an agent is 'harmed' by a response when and only when that response decreases the agent's well-being, or prohibits that agent from an increase in well-being that s/he would have otherwise obtained. Those who reject the assumption are welcome to consider my use of 'harm' as a mere expedient for the more cumbersome disjunction in the analysans.

[209] Sher 2006, 130. See also Zimmerman 1988, 152; Watson 1996; Arpaly 2006, 9-39; Shoemaker 2011; Nelkin 2011, 34-35

prescribed medication with his mild alcohol consumption. He would certainly have resisted his motivation otherwise.

Janus is responsible for this betrayal in more than just a causal sense. He's gossipy. That's partly why he betrayed you. So it is apt to morally disapprove of him. The aptness of this moral disapproval also justifies a few sorts of responses. First, you're justified in altering any relationship practice which rests on prior beliefs that Janus's attributability defeats. For example, your friendship with Janus was probably based, like many close friendships, on a high level of trust of which you now have justification to believe him less capable. You should adjust your relationship accordingly.[210] Second, Janus is 'answerable' for his betrayal in that you're justified to demand that, insofar as he is able, he explain his reasons for it, and that he justify or excuse it if he can.[211]

It is understandable that you would also privately feel blaming emotions towards Janus, and be prepared to target him with harmful blaming responses (detailed below), should he fail to provide a viable justification or excuse for his behavior. I take this private form of blame towards attributable agents to be morally permissible, since it doesn't affect anyone other than the blamer. It is worth noting, however, that it is extremely difficult to keep blame private. It typically either manifests in some harmful action, or dissipates. In this case there is apparently no positive reason for the former.

According to Janus, his betrayal was in large part a result of forces beyond his control. He simply could not have foreseen that his medication would loosen his lips in

---

[210] Scanlon 2008, 128-31; Shoemaker 2011.

[211] David Shoemaker (2011 and 2015) argues that answerability is a form of responsibility distinct from attributability and liability. Angela Smith (2012) disagrees. My arguments are consistent with either conclusion.

the way that it did (there was no warning on the label, his prior prescription didn't have this effect, etc). If that's true, it constitutes a reasonable excuse for his behavior. So although it is apt to morally disapprove of Janus, to adjust your relationship practices, demand him to answer for himself, and permissible to privately feel benign blaming emotions towards him, he is not *liable* to any response that will further harm him, including being targeted by blame, or to agents being disposed to target him with blame. He is not, as Watson puts it, 'accountable' for his actions.[212]

This is what distinguishes liability from attributability. An agent can be morally responsible for doing wrong in an attributability sense even if s/he isn't liable to any responses that harm in ways beyond those brought about by mere relationship adjustment or demands for justification. But an agent is morally responsible for doing wrong in a liability sense only if s/he is liable to such a response for it.

1.3 Desert and liability

There are various moral considerations that might render a transgressor liable to harmful responses such as blame and punishment. That the agent 'deserves' such responses is the consideration most often implicated by ascriptions of moral responsibility. When an agent deserves some potentially harmful response, s/he deserves it *just* because s/he transgressed in the way s/he did, and not for any further

[212] Watson 1996. Also see Shoemaker 2015, 113. I prefer to follow Karin Boxer (2013, 3) in using H.L.A Hart's term 'liability' because 'accountability' suggests an interpersonal nature, which I deny. See especially section 2.2 below.

reasons, such as those generated by consequentialist or contractualist theories.[213] For example, suppose that

> *Janus's liability*: you discover Janus's betrayal was not an unforeseeable accident. He's not on any medication. The alcohol didn't impair his judgment. Though he likes you, and thinks of you as a friend, he's been jealous of your charm and charisma for quite some time. So it turns out that he freely and knowingly divulged your secret in an attempt to bolster his own social standing at your expense. And then he lied about it.

Janus deserves some potentially harmful response for this transgression, blame being an obvious choice. Such a response may bring about better states of affairs. It may be that Janus tacitly agreed, or would agree, to abide by rules which, if broken, make him liable to such responses. Such a response may even be inevitable, given your anger at Janus. But these reasons to potentially harm Janus do not seem to capture the basis of his desert. Janus deserves some potentially harmful response regardless of any (hypothetical or actual) agreement, or the value of the states of affairs that result. He deserves it *just* because he freely and knowingly divulged your secret for nefarious reasons.

This is not to say that anyone ought to harmfully respond to Janus, all things considered. If for example a peaceful response to Janus guarantees a far better state of affairs (Janus's reform, your relationships repaired, your friends putting the secret out of

---

[213] This 'basic' desert formulation is Derk Pereboom's (2014, 2). Also see Scanlon 2013, 101- 102; Feinberg 1970, 55-94; Clarke 2005, 21; Strawson, G. 2002, 452; Waller 2011, 2-5; McKenna 2012, 150.

their minds, etc), then although Janus deserves a potentially harmful response, it may be that no one ought to respond to him that way, all things considered. This is because, when an agent deserves some response, s/he is only *defeasibly* liable to the response in one of the following ways[214]:

> *Goodness*: there is some non-instrumental good in the agent receiving the response.[215]
>
> *Reason*: there is a defeasible non-instrumental reason to respond to the agent in that way.[216]

I haven't the space to adjudicate which interpretation is correct. I'll just assume that, for any agent and allegedly deserved response, showing that both the goodness and reason interpretation apply is sufficient to show that the agent deserves the response. I'll assume that showing neither interpretation applies is sufficient to show that the agent does not deserve it.

Some may object that there are alternative forms of liability for wrongdoing that do not involve desert, as defined.[217] I address alternative forms of liability in the final section. But I should note that my focus in this paper is on the sizable number of

---

[214] Kleinig 1971, 76; Hanna 2013.

[215] Ross 1930, 134-138; Miller 1999, 136; Sher 1987, 194-198; Hurka 2001, 8-12; Moriarty 2003, 520; Bennett 2002, 147; Kagan 2003, 93; McKenna 2012, 172; Pereboom 2014.

[216] Feinberg 1970, 60; Zimmerman 1988, 162; Schmitdz 2002, 774. Some (Mundle 1954, 217 and Moore 1993, 15) claim that the reason is always *sufficient*. Others (Murphy 1971 and 1973; McLeod 1999, 193) argue that the reason comes in the form of an obligation.

[217] For a consequentialist theory of liability, see Smart 1961. For a discussion of other alternatives, see Boxer 2013, 3-5.

philosophers who agree with me that desert, so defined, is the justification at stake in cases of liability for wrongdoing. My concern is to instead challenge the sort of response that is alleged to always be at issue. Specifically, I'm taking aim at those who agree with my examination thus far, and endorse

> *The blame-focused view*: An agent is morally responsible in a liability sense for doing wrong just in case s/he deserves to be blamed for it.

This includes many of those working in the free will debate.[218] Those who reject desert as the justification at issue can take the ensuing arguments to be aimed at a more cautious conclusion; *if* desert, so defined, is at issue in cases of liability for wrongdoing, then it is the desert of intrapersonal responses like guilt rather than interpersonal responses like blame that explains the wrongdoer's liability. As mentioned in the introduction, I first plan to argue for this conclusion by establishing that an agent deserves guilt whenever s/he deserves a harmful interpersonal response such as blame. A better understanding of blaming dispositions propels us towards this conclusion.

1.5 The harms of deserved blame

As Christopher Bennett, Michael McKenna, and Karin Boxer have recently pointed out, blame is constituted by three sorts of dispositions, each of which tends to harm the blamed. First, anger disposes blamers to purposely *alienate* the blamed from

---

[218] See especially Fisher 2006, 62; McKenna 2012, 150; Pereboom 2014, 2.

the moral community. As Bennett puts it, when you blame an individual "the first thing you are disposed to do is to stop speaking to them, and to stop helping them. You stop offering them the basic goodwill that human beings might normally expect of one another."[219] Michael McKenna points out that such alienation is a deprivation of what Feinberg calls 'welfare interests'. Specifically, it deprives the blamed of "the capacity to engage normally in social intercourse and enjoy and maintain friendships."[220] Typically, this is harmful to the blamed agent.

Second, blaming anger disposes one to *seek repentance* from the blamed.[221] Bennett and Boxer point out that blamers desire that repentance come by way of some personal sacrifice.[222] I suspect that this desire is aimed more generally at making the blamed suffer, a desire that many take to be *sine qua non* of anger.[223] Strawsonians such as McKenna, Boxer, and Bennett,[224] on the other hand, take this desire to be constructive, aimed at motivating the blamed to repay the victim, and more importantly to "resuscitate the credibility of" his social commitments, and in general "recover his place in the moral community".[225] Perhaps there are ways of accomplishing these goals

---

[219] Bennett 2002, 149-152.

[220] Feinberg 1984, 37; McKenna 2012, 136.

[221] Bennett 2002, 151; McKenna 2012, 138-141; Boxer 2013, 125.

[222] Bennett 2002, 159; Boxer 2013, 127-133.

[223] Vidmar 2001, 43; Darley and Pittman 2003, 333-334.

[224] Strawsonians are those who claim that moral responsibility is not solely a fact about an agent, but rather is generated by the natural attitudes, specifically blaming attitudes, that are sometimes elicited by that agent's interpersonal interactions and relationships with others (Strawson 1982). A powerful critique of Strawson's account is that giving-in to naturally elicited blaming attitudes can sometimes, perhaps always, mean treating individuals in a way that they do not deserve. Bennett, Boxer, and McKenna are among the minority of Strawsonians who argue that Strawson can respond to this critique by embracing desert.

[225] Bennett 2002, 160-161.

without the blamed agent suffering in any way. But that seems difficult. As McKenna acknowledges, seeking repentance involves interfering in the blamed agent's life, which is likely to cause substantial emotional disruption.[226]

It is intuitive that a blameworthy transgressor like Janus is liable to some degree of the alienation and repentance that blame tends to bring about. But alienation is instrumental to, and repentance arguably requires, what I take to be the more fundamental harm that blame is intended to elicit. Most agree that blame is a form of 'guilting' in that blamers intend their blame to result in the blamed feeling *guilty* for transgressing.[227] This intentional guilting is not without a moral purpose. Those working on moral responsibility urge that a blameworthy transgressor gets all that s/he deserves only if s/he feels guilty for the transgression. Consider the following variation of a thought experiment by Randolph Clarke.

> *Janus's guiltworthiness*: You overtly blame Janus for his transgression. In doing so, you alienate him. You communicate to him that you desire, even expect, his repentance. He tells you, sincerely, that he already shares your moral disapproval, and has already begun reforming himself. Over the next days and weeks, he shows this to be true, all the while trying to make amends, telling you that he deeply values and respects you as a person, and that he wishes he had

[226] McKenna 2012, 138-139.

[227] Gibbard 1990; Bennett 2002, 152-153; Darwall 2006; Smith 2007, 477; Shoemaker 2007, 91; McKenna 2012, 139-140; Boxer 2013, 125.

> acted otherwise. He also apologizes profusely. However, he candidly admits during his apology, "I haven't felt even a little guilty for divulging your secret."[228]

This emotional failure is jolting. Janus clearly understands that he's wronged you, shares your disapproval, and is responding to every legitimate moral demand. But his candid admission that he feels no guilt whatsoever for wronging you generates the strong intuition that Janus has failed to receive all that he deserves for his transgression. It is not the lack of guilt *per se* that explains the intuition. For it would persist even if we stipulated that Janus felt guilty for his transgression for purely non-moral reasons; for example because in transgressing he was imprudent, irrational, or careless in execution. The sort of guilt that explains why Janus does not get all that he deserves is *moral* in nature.[229] By this, I mean

> *Moral guilt*: sorrow in virtue of morally disapproving of oneself.

Moral guilt (henceforth simply 'guilt') is the potential harm of blame that Janus must receive in order to get all that he deserves.[230]

1.6 Blameworthiness implies guiltworthiness.

---

[228] Clarke 2013, 156.

[229] I take it that an emotion is 'moral' whenever it is by virtue of a belief that there has been a moral transgression of some kind, a belief implicated by moral disapproval (Wallace 1994, 33-38).

[230] Bennett 2002, 151; Darwall 2006, 71, 168; Shoemaker 2007; McKenna 2012, 138-141; Boxer 2013, 125; Clarke 2013, 157; Rosen 2015, 82-83.

I think this is because a blameworthy agent like Janus *deserves* to feel guilt for the transgression.[231] More generally, blameworthiness implies guiltworthiness. Several considerations support this thesis. First, it is intuitive in two distinct ways. From the perspective of an accuser, it is intuitive that there is some non-instrumental good in Janus feeling the guilt that he lacks, and that he has a non-instrumental reason to see to it that he feels guilt. The non-instrumental goodness of his guilt, and the defeasible reason for him to bring it about, seem to obtain *just* because he freely and knowingly transgressed. This other-focused intuition is evidence that blameworthiness implies guiltworthiness.

We can also judge Janus from a first-personal perspective, as if in his shoes. This is rarely a perspective that is adopted when considering a transgressor's deserts. But as Michael S. Moore has pointed out, when we put ourselves in a transgressor's position, we realize that we ourselves would feel guilt.[232] I take it that we would almost certainly see our guilt as having some non-instrumental goodness, and ourselves as having a reason to bring it about. Moore points out that, all else being equal, a person's judgement that s/he would deserve some response in a transgressor's position is good evidence that the transgressor deserves that response.[233] This self-focused intuition is further evidence that Janus deserves to feel guilt for his transgression, and that in general blameworthiness implies guiltworthiness.

---

[231] Clarke 2013; Rosen 2015, 82-83.

[232] Moore 1993, 26.

[233] Moore 1993, 27.

This claim is not merely intuitive. It is also the best explanation for the compelling claim that Janus *must* feel guilt in order to get all that he deserves. One might try to explain why Janus must feel guilt in order to get all that he deserves by claiming that

> *G*: any given blameworthy transgressor must feel guilt in order to get all that s/he deserves, even though s/he does not deserve the guilt itself.[234]

G does not account for our intuitions that guilt is deserved. But more worryingly, given that a primary purpose of blame is to elicit guilt, *G* implies the following: just because an agent transgresses, there is some non-instrumental good, or a defeasible reason, to blame her in order to elicit her guilt, even though there is *no* non-instrumental good, and *no* defeasible reason to see to it, that the transgressor feels this guilt. That is not an outright contradiction. But it is not *prima facie* plausible when compared to the thesis that blameworthiness implies guiltworthiness. So the onus is on the defender of *G* to offer a compelling argument in its favor.

One way to argue for *G* is to argue that a blameworthy transgressor must feel guilt for some reason that is not desert-based. For example, some might argue that moral guilt is made apt by mere attributability norms. However, I take it that guilt necessarily pains the agent who feels it, and is therefore harmful, all else being equal. Those norms "that require that we believe propositions that are true and that we accept moral principles that are justified" can't alone offer a reason or make it good to feel pain. A normative justification like desert is clearly at play.

---

[234] Boxer 2013, 10, 62-63, 127-133.

Granted, it is difficult (though I do not think impossible) to morally disapprove of oneself - specifically to desire for the right moral reasons that one hadn't acted wrongly - without feeling the pain of guilt. But this feeling isn't merely a regrettable concomitant of apt disapproval. For if it were, we should have no moral qualms with the use of some emotional analgesic to relieve it. I doubt this use would be seen as morally appropriate. A blameworthy wrongdoer who tries to rid themselves of the pain of guilt is not ridding themselves of a feeling morally equivalent to a headache. They're failing to emotionally respond as they ought.

Some might argue instead that guilt is merely required for a wrongdoer to recognize that what s/he did is morally bad or wrong, and that this recognition is in turn required for a blameworthy agent to get all that s/he deserves, even though s/he doesn't deserve the guilt itself.[235] In my view, feeling guilt is not required for recognition of wrongdoing. For example, I believe (due to the work of Alaistar Norcross) that supporting the factory farm industry through consumer purchases is wrong. But my occasional enjoyment of factory farmed meat preempts any guilt that I might otherwise feel when recognizing this fact. Moral emotions are too influenced by our morally imperfect values to align with our recognition of wrongdoing. Sometimes we recognize that an action is wrong, but guiltlessly do it anyway.[236]

Rather, feeling guilt is part of what it is for a wrongdoer to genuinely *care* about the moral norm that s/he's transgressed. Janus simply doesn't care enough about his betrayal. I don't care enough about my wrongful meat consumption. Hence the lack of

---

[235] This is one way to interpret Darwall 2006, 71, 112, 168; 2009, 144. Also See Pereboom 2017.

[236] Clarke 2016, 126-127.

guilt. It is this fact that I think helps explain why blameworthiness implies guiltworthiness. A morally decent person is one whose concern about moral norms brings about the feeling of guilt when s/he's freely, knowingly, and inexcusably done wrong. Each of us has a non-instrumental reason, and it is non-instrumentally good, to be a morally decent person. So each of us has a non-instrumental reason, and it is non-instrumentally good, to feel guilty when we've freely, knowingly, and inexcusably done wrong. To blame someone is to attempt to get them to respond to this justification. *G* is, then, implausible. Blameworthiness implies guiltworthiness.

1.7 Guilt vs blame

Some might think that guilt is just a species of blame.[237] If that is right, then the thesis that blameworthiness implies guiltworthiness could be seen as rather trivial; showing nothing more than that transgressors who deserve blame from others also deserve blame from themselves. There is no denying that some instances of guilt are colloquially classified as a kind of 'self-blame'. We often tell those who are unjustifiably sorrowful in virtue of their moral disapproval to 'stop blaming themselves.' However, even if guilt is a species of self-blame, the conclusion that blameworthiness implies guiltworthiness is, as we will see throughout the next section, far from trivial. It is the first step in an argument leading us to the conclusion that it is always a self-directed emotion (whether we call it 'guilt' or 'self-blame') rather than an other-directed emotion that explains moral responsibility for transgressions.

---

[237] This is another way to interpret Darwall 2006, 168; 2009, 144. Also see Morris 1988; 66-67; Wallace 1994, 66-67; Skorupski 1999, 142; McKenna 2012, 72-74.

Still, it is worth briefly noting why it is a mistake to classify guilt as a kind of blame in anything more than a colloquial sense. Although there are cases in which we 'blame' ourselves, I take paradigm examples of self-blame to lack feelings of sorrow. Self-blame is instead constituted by hostility, frustration, and even self-loathing. As discussed above, these angry feelings dispose us to harm in various ways. One of them is to alienate the target of the emotion. We cannot literally alienate ourselves from ourselves.[238] But we can, and I think often do when self-blaming, see our transgressions as manifested in a version of ourselves that we do not fully identify with. To self-blame is to distance ourselves from that morally imperfect manifestation, in a way kicking ourselves for failing to express what we take to be our true moral character.

Such an attitude might be a good form of moral self-regulation. And it's a way of expressing a commitment to live up to a moral ideal. But it's rarely true or epistemically justified in regards to transgressions for which we are morally responsible. Moral deficiencies endure. It takes time and effort to change them. When we're being honest with ourselves, we rightly see those transgressions we are morally responsible for as indicating who we are as persons. To avoid the aforementioned alienation and instead honestly identify ourselves as immoral, here and now, elicits forms of sorrow; remorse, disappointment, sadness, and the like. Such an emotion nearly always obtains in the absence of any angry feelings or dispositions, or even the belief that anger towards oneself is deserved.[239] Guilt doesn't aim to hurt. It just hurts. I think that's enough to warrant a distinction in kind between guilt and self-blame.

---

[238] Thanks to Michael McKenna for making this point, and for helpful discussion on this issue.

[239] Waller 1990, 165-166; Pereboom 2014, 186-187.

2 Guiltworthiness doesn't imply blameworthiness.

Blameworthiness implies guiltworthiness. So guilt is a genuinely deserved response in all of those cases in which blame is deserved. This shows that it is at least a candidate for explaining the liability sense of moral responsibility for transgressions in all of the cases in which blame is alleged to explain moral responsibility.[240] In this section, I offer a few kinds of cases supporting the conclusion that morally responsible transgressors can lack the desert of blame. However, in each case, the morally responsible agent deserves guilt. So blameworthiness for transgressions is more fragile than moral responsibility for transgressions, while guiltworthiness is not. So the guilt-focused view is a viable alternative to the blame-focused view, one that I also argue offers solutions to puzzling features about moral responsibility.

2.1 The moral community

According to many advocates of the blame-focused view, it is some fact about the value of moral-interpersonal relations between members of the moral community that explains any given agent's blameworthiness for a transgression. As a test case for this claim, Michael McKenna has us suppose that

> There is a world where a solitary person, call him Robinson, mercilessly beats his dog. There is no moral community, no practices of holding morally responsible built up out of the broader array of adult interpersonal relations Strawson wrote

[240] I take it that this is one reason why Gibbard (1990, 42) claims that it is both blameworthiness and guiltworthiness that explains moral responsibility.

> about, no conventions for blaming, nor for offering excuses or justifications, and so on. No...moral responsibility practices. There is just brute nature all around, this one person, Robinson, and his cruelty to his pet....Robinson can apply moral predicates, and so can understand that his beating Rover is cruel. Perhaps he thinks it is morally bad and believes what he is doing is morally wrong. Nevertheless, [he] has no concept of an interpersonal moral responsibility exchange...no idea what it would mean to be blamed, to be held to account for his actions, for him to offer apology to others for his conduct...[241]

Here we're imagining an agent who's circumstances make it virtually impossible for him to recognize or care about anyone's moral authority. As David Shoemaker remarks of a similar case, Robinson seems to be the sort of agent who "accepts that he has reason to refrain from [doing wrong but] denies that that reason has its source in anyone's authority as a member of the moral community to demand it."[242] Robinson's nature precludes his ability to "take himself to be responsible *to us*."[243] So, conclude McKenna and Shoemaker, individuals like Robinson are not blameworthy.[244] There is neither non-instrumental goodness in, nor a non-instrumental reason for, blaming Robinson.

I concur. As Stephen Darwall points out, blameworthiness is essentially interpersonal. It is a form of moral address the purpose of which is the communication of *second-personal reasons*. In general, these are reasons that an agent has due to

---

[241] McKenna 2012, 107-109.

[242] Shoemaker 2007, 86.

[243] Shoemaker 2007, 87.

[244] Shoemaker 2007, 87; McKenna 2012, 109.

"some valid claim or demand of some [particular individual or group of individuals] having practical authority with respect to the agent and with which the agent is thereby accountable for complying."[245] The specific second-personal reason relevant here is the *basic demand*: a demand that members of the moral community maintain and evince a certain quality of will towards fellow community members. Darwall and others argue that an agent deserves blame *only* if the agent can understand and apply this basic demand. Robinson can act in accordance with the basic demand, but cannot understand and apply it precisely because his capacity to see others as moral authorities is altogether absent.[246] So Robinson is not blameworthy. I take it that the same goes for any interpersonal response alleged to be deserved by a wrongdoer. In blaming, guilting, punishing, etc, we're trying to communicate the way in which the agent is accountable for complying. These messages are alien to Robinson. So he does not deserve any interpersonal response.

If the blame-focused view is correct, this fact entails that Robinson is not morally responsible for beating his dog. Indeed, McKenna, Darwall, Shoemaker and others argue that agents who are insensitive to the basic demand also can't be liable for their wrongdoing.[247] This conclusion is puzzling. Although Robinson does not understand the interpersonal features of moral reasons and moral responsibility practices, he clearly recognizes the *im*personal form of such prohibitions such as the prohibition against

---

[245] Darwall 2009, 143.

[246] Shoemaker 2007, 86.

[247] Darwall 2006 and 2009; Shoemaker 2007; McKenna 2013, 107-110. The view originates with Strawson 2008, 29.

causing pain for no good reason.[248] As McKenna makes sure to point out, Robinson "can apply moral predicates, and so can understand that his beating Rover is cruel. Perhaps he thinks it is morally bad and believes what he is doing is morally wrong." Robinson freely beats his poor dog anyway.

Robinson's morally impoverished upbringing might make it the case that he is now wholly incapable of feeling guilt or any other negative self directed emotion. If so, then I agree that he is not morally responsible for it in a liability sense. After all, in that case there would be no harmful response that would be intelligible to Robinson as anything other than retaliation. There would be no way for him to *care* about morality in anything more than the cold and detached ways of a severe sociopath.

However, it is difficult to conceive of Robinson having even a minimal understanding of obligations without the capacity to feel guilt.[249] It's far more conceivable, and in any case we can stipulate, that Robinson indeed has the capacity to feel guilt for beating his dog. Assuming that's right, it's intuitive that Robinson's moral sensitivity is enough to make the feeling of guilt a deserved response. How much guilt Robinson deserves might be mitigated by his impoverished moral understanding. But to say that he's not liable to a harmful response - full stop - is too much of a stretch. Interpersonal incompetence precludes the desert of interpersonal responses, not liability.

One objection to this conclusion is that Robinson did not do wrong because ϕ-ing is wrong only if agents who freely, knowingly, and inexcusably ϕ are blameworthy for ϕ-

---

[248] Shoemaker 2007, 90.

[249] Though for a critique see Harman 2009.

ing.[250] I don't wish to reject the view that the desert of some form of potentially harmful response for ϕ-ing is necessary for the wrongness of ϕ-ing. But I fail to see why blame is the only response that qualifies. Surely guilt is also among these potentially harmful responses. Consider Mill's claim that "we do not call anything wrong unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, by the opinion of his fellow creatures; if not by opinion, *by the reproaches of his own conscience*."[251] One reading of the passage is that guiltworthiness for ϕ-ing is the only desert necessary for the wrongness of ϕ-ing. Kurt Baier endorses a kind of Millian view, arguing that an action is wrong if and only if it "licenses interference" when committed.[252] By "interference", Baier has in mind "pressures which are in themselves obnoxious...methods such as handcuffing, jailing, fining, and possibly simply 'condemning'."[253] I see no reason why one's own guilt could not also qualify as 'obnoxious' in Baier's sense.

Another objection to the conclusion that Robinson is blameless but morally responsible relies on the following two premises: (1) Any wrong action is made wrong by second-personal reasons against it.[254] (2) An agent knowingly does wrong only if s/he understands what it is that makes the action wrong. It follows from (1) and (2) that an agent knowingly does wrong only if s/he understands second-personal reasons. So, because Robinson doesn't understand second-personal reasons, he simply can't

---

[250] Gibbard 1990, 42; Skorupski 1999, 142.

[251] Mill 2009, ch.5, 88

[252] Baier 1966, 223-226.

[253] Baier 1966, 223.

[254] Darwall 2006, 91-104, 2009.

knowingly do wrong. But it's a given that an agent is morally responsible for doing wrong only if s/he *knowingly* does wrong. So Robinson can't be morally responsible for doing wrong.

The epistemic restriction in (2) is too stringent, especially under the assumption that (1) is true. Just as a person can know that their gambling is irrational, even if s/he's incapable of understanding the fundamental features of irrationality, a person can know that s/he's doing wrong even if s/he's incapable of understanding the fundamental features of morality. For example, no one has a convincing explanation for why it is wrong to bring about lives barely worth living that would not have existed otherwise. But that doesn't excuse those who do. There's just something wrong about it. That's plain enough.

Advocates of the blame-focused view may concede that guiltworthiness explains the moral responsibility of those wrongdoers, like Robinson, who exist outside of the moral community. But they may argue that blameworthiness at least explains the moral responsibility of those wrongdoers who are less interpersonally incompetent. Responsibility for our actions, they may argue, is essentially interpersonal for members of the moral community. There are several considerations that militate against this claim.

2.2 The suberogatory

Suppose that

> Lisa is exiting a public place. A stranger behind her has his hands full. But Lisa's mood is a bit sour. So she flippantly decides not to open the door for him, even though she knows that this guarantees him a small amount of struggle.

Most of us intuit that Lisa is morally responsible for her action in more than a mere attributability sense. After all, to fail to prevent another person a bit of struggle without a good reason is not just poor etiquette. It is *morally* bad, evincing a lack of good will on Lisa's part. Lisa knows this. And she freely did it anyway. That is grounds for more than mere moral disapproval. It is also grounds for guilt.

Suppose, for example, that Lisa knows that she will feel a small amount of guilt if she looks back to see the man struggle with the door, and won't feel guilt otherwise. Her not looking back seems morally deficient. It seems better, and she seems to have more reason, to choose the former option. And wouldn't you, in the same scenario, judge there to be a reason, and judge it to be non-instrumentally good, to elicit within yourself a small amount of guilt for your action? I suspect an affirmative answer runs deep for many of us. That's indication that Lisa deserves guilt for her action.

Suppose further that Lisa's mother witnessed the scene while waiting in the car. She justifiably suspects that bringing Lisa's attention to the man's struggle will make Lisa feel guilty for not helping the man. She does so, saying flatly "look at that poor man Lisa. He needed help." She may even ask, in a non-hostile tone, "why didn't you help him?" Such non-hostile 'guilting' behavior is quite common. But there are general moral prohibitions against bringing about easily avoidable harms for no good reason. So Lisa's

mother needs a justification for guilting Lisa in this way. All else being equal, the fact that Lisa *deserves* to feel guilty seems an excellent justification.

However, it strikes me as counterintuitive that Lisa deserves to be blamed in such a case. Suppose that Lisa's mother not only attempts to bring about Lisa's guilt, but does so in a blaming manner by attempting to alienate her and seek her repentance. She may seek repentance by, for example, communicating to Lisa in an angry tone that she wants Lisa to go apologize to the man. And she may alienate Lisa by remaining disapprovingly silent throughout some portion of the car ride home. Any degree of these blaming behaviors seems excessive. Although the man probably thought Lisa's action to be bad, it is unlikely that he 'held it against her' in a way that would make an apology anything more to him than a further inconvenience. Anger at Lisa solely because of her action is extreme. Lisa's action is morally bad. But it's not *that* morally bad. Lisa is guiltworthy but not blameworthy.

That Lisa is not blameworthy is also entailed by a commonly-held restriction on blameworthiness. Although Lisa's action is morally bad, it is also arguably morally permissible, best described as a suberogation.[255] Many argue, correctly I think, that *blameworthiness requires wrongdoing:*

> *BRW*: an agent deserves blame for ϕ-ing only if it is wrong that s/he ϕ's all-things-considered.[256]

---

[255] Driver 1992 and Macnamara 2012.

[256] Widerker 1991, 223; Wallace 1994, 12, 135; Copp 1997, 445; Pereboom 2009, 170; Nelkin 2011, 100-107; Boxer 2013, 64. I intend ϕ-ing to include refraining.

One reason to think that *BRW* is true is the close relation between obligations and demands. Derk Pereboom, citing Dana Nelkin, contends that "'ought' propositions that specify what an agent [is morally obligated] to do are essentially action-directed, so that if '*S* ought not do *A*' is true, then as a matter of the meaning of 'ought' propositions, or of the essential nature of obligation, *S* is thereby directed to refraining from *A*" by way of a demand.[257]

This being so, I take it that the fairness of demands are linked to moral obligations in the following way. It is fair to an agent to demand her to not ϕ only if it is wrong that s/he ϕ's all-things-considered.[258] To demand that an agent not ϕ is to request that the agent not ϕ, and to sincerely indicate that one will repudiate her for noncompliance with this request. That is why to repudiate an individual for acting permissibly is also unfair to her. It is akin to saying 'I do not accept the acceptable'. The locution is not contradictory. But it does not fit comfortably in the mouth. Blaming behaviors - specifically alienation and the seeking of repentance - are widely understood as unequivocally symbolizing repudiation.[259] They are ways of indicating, at a minimum, that the blamer does not accept some action or characteristic. This is one reason that *BRW* is appealing. Blaming an agent for doing nothing wrong is, at least

---

[257] Pereboom 2014, 139. See Nelkin 2011, 108-116.

[258] Note that this principle does not entail, nor do I mean to suggest, that it is never fair to others or even right, all things considered, to demand A to ϕ if it is not the case that A ought to ϕ. Cases of eminent domain, for example, seem to be cases in which it is unfair to an agent to demand that she forfeit her property to the state precisely because it is morally permissible for her to try to retain that property. Nevertheless, in some of these cases it is arguably both fair to others and right, all things considered, to demand her to do so.

[259] Feinberg 1965; Darwall 2006; Boxer 2013.

under current conventions, to repudiate an acceptable action. That is unfair. And that unfairness is sufficient to undermine any allegedly deserved blame.

However, *BRW* leaves it open to us that Lisa is guiltworthy for her suberogation. This is as it should be. If ϕ-ing is morally permissible but still morally bad, then although it is unfair to an agent to demand that she refrain from ϕ-ing, it may still be fair to the agent to strongly recommend that she not ϕ.[260] Such a recommendation against ϕ-ing neither indicates nor requires the fairness of repudiation. Guilt and non-emotional forms of guilting are typically neither interpreted nor intended to be forms of repudiation. So for an agent to feel or be asked to feel guilt when she freely and knowingly suberogates is not always unfair to that agent. Moreover, the norms governing negative responses to oneself are not constrained by what *others* can legitimately expect or demand. Those norms are intrapersonal. Lisa is guiltworthy for her suberogation, but also blameless for it.

The view that those who suberogate (suberogators?) can be guiltworthy but not blameworthy offers a solution to a puzzling feature of moral responsibility. Consider the following three propositions.

(a) ϕ-ing is suberogatory.

(b) *A* is morally responsible in a liability sense for ϕ-ing.

(c) *A* is *not* blameworthy for ϕ-ing.

---

[260] Pereboom 2014, 140-141.

Cases abound in which each of these propositions seem true.[261] If the blame-focused view is right, then an agent who is not blameworthy simply *cannot* be morally responsible. So if the blame-focused view is right, (b) and (c) cannot simultaneously be true. This poses the following problem. For cases like Lisa's, those who endorse the blame-focused view must either deny (b) in an attempt to affirm (c), or deny (c) in attempt to affirm (b). Denying either is, as I have been urging, independently implausible. But as I pointed out above, *BRW* and (a) entail that the suberogator *can't* be blameworthy. So those who deny (c) must also commit themselves to further implausibilities. They must also deny *BRW*,[262] or deny (a).[263] The latter denial is not plausible, at least in cases like Lisa's. And as I argue in the appendix to this paper, arguments against *BRW* are unsound.

The guilt-focused view allows us to avoid such inconsistencies. If the guilt-focused view is right, an agent can be morally responsible even if s/he isn't blameworthy. So the guilt-focused view allows the consistency of (b) and (c). Moreover, since *BRW* does not preclude guiltworthiness for suberogations, the guilt-focused view also allows the consistency of (a) and *BRW*. So the guilt-focused view allows us to consistently affirm (a), (b), (c), and *BRW*. The availability of such a simple and intuitive way to retain the consistency of independently plausible claims is an attractive feature of the view. It is also a reason to accept it.

2.3 The instrumental character of blame

---

[261] For similar cases see Driver 1992 and Macnamara 2012.

[262] Driver 1992; Macnamara 2012, 145; McKenna 2012, 182-187.

[263] Nelkin 2012, 104-107.

Unlike Robinson, Lisa is perfectly morally and interpersonally competent. But like Robinson, she is not blameworthy, despite being morally responsible for her transgression. So advocates of the blame-focused view cannot claim that blameworthiness always explains the moral responsibility of fully competent transgressors.

Advocates of the blame-focused view may point to the fact that a feature these cases share is that neither transgressor can be legitimately targeted with the demands that come along with blame. Advocates may then claim that guiltworthiness explains the moral responsibility of such transgressors, but that blameworthiness explains the moral responsibility of anyone who is legitimately targeted with demands; that is, wrongdoers who are members of the moral community. Those (like Lisa) who merely suberogate or (like Robinson) do not understand second-personal reasons are in a sense only 'partly' responsible for their wrongdoing by way of their guiltworthiness. Because they don't have to answer to anyone, and hence aren't blameworthy, they aren't 'fully' morally responsible. This response is challenged by two features of allegedly deserved blame, both of which conflict with our intuitions about the nature of deserved objects.

If the blame-focused view is right, a blameworthy individual deserves *others* to target her with blame. This is puzzling. In so many cases, an agent's receipt of her desert affects her necessarily, others only contingently. We routinely say, for example, that hard workers deserve the fruits of their labor, that the most qualified deserves to get the job, and that the fastest runner deserves to win the race. Most pertinently, consider desert for morally good actions. Suppose Sadam anonymously donates his kidney in an attempt to save a child's life. The child survives. This causes Sadam joy. Intuitively,

Sadam deserves his joy. No positive interpersonal response, for example praise, is necessary for him to receive this desert. More generally, objects of desert needn't affect anyone but the deserving agent. Each agent's desert is theirs and theirs alone.

Blame necessarily affects the blamer but, as in the Robinson case, and the example below, only contingently affects the target. So the blame-focused view posits an object of desert that is fundamentally different from the objects of desert cited above. It is one that affects the blamed contingently, blamers necessarily. This implication is especially troubling because blaming emotions - anger, irritation, frustration, and the like - tend to affect blamers harmfully. They can distort our beliefs, become dangerously misplaced, escalate into altercations, strongly incline us to act irrationally, or expose us to frustration and disappointment when the target of blame doesn't conform to expectations.[264] Being angry, irritated, or frustrated also tends to feel bad.

This is why blame has always struck me as instrumental to desert, and never deserved *per se*. It is something that desert justifies only when blame achieves its primary aims. Guilt is a primary aim of blame. It affects the guilty necessarily. And the non-deserving needn't be affected at all in order for a wrongdoer to feel guilt, much less that they suffer any harm. The guilt-focused view certainly leaves it as a possibility that blame is sometimes deserved. But it avoids the above difficulty by allowing that blame is something desert justifies only when it is a means to some amount of deserved guilt. For example, return to the case of Janus. Why didn't he feel guilty for wronging you? Suppose you discover that

---

[264] For a nice overview of some of these effects, see Goldberg, Lerner, and Tetlock 1999, 781-783.

Janus's blamelessness: Janus was raised in a hostile household with parents and siblings who were constantly looking to target one another with anger. Because they were generally cruel, becoming emotional in response to their hostility would only attract relentless teasing and bullying. Janus became severely traumatized by this at a young age. He survived by instinctually shutting his emotions down in response to even the mildest indication of anger. This trauma never left him. As a result, any indication of anger makes Janus emotionally numb. He can't help it. It's just the way he's been conditioned. However, Janus is perfectly capable of feeling guilty for wronging you. In fact, he would have already felt immensely guilty if, instead of blaming him, you had turned the other cheek and calmly described to him, in a non-blaming manner, your moral disapproval, and the embarrassment and hurt you felt when he wronged you.

Janus is unusual. But he's by no means an unknown type. Some people are too traumatized by circumstances beyond their control to respond normally to blame. Some become defensive and hostile. Others, like Janus, just shut down. The only emotional effect blame has on him is to inhibit guilt that alternative responses would have elicited. Granted, blaming him expresses legitimate moral demands. But these moral demands can just as easily be communicated through a variety of non-hostile responses. Given his emotional difficulties, Janus is more receptive to moral demands conveyed in a non-blaming manner. Moreover, recall that one of the details of the case is that Janus morally disapproved of himself and began responding to the demands you eventually

made of him (to reform, apologize, repent, etc) without anyone needing to first communicate those demands to him. Moral demands are like that. They 'demand' by way of normative, not verbal, force. That being so, I find it strange to think that there is any non-instrumental goodness in, or non-instrumental reason for, blaming Janus *just* because he did wrong. That's because blame doesn't serve it's primary purpose in regards to desert. Specifically, blame won't bring about the guilt that Janus clearly deserves.

If that's right - Janus doesn't deserve to be blamed - then the blame-focused view entails that Janus is not morally responsible (in a liability sense) for wronging you. But it's our strong intuition that Janus *is* morally responsible for wronging you. That's the answer that the guilt-focused view predicts. Janus is morally responsible for wronging you because he deserves to feel moral guilt for doing so. This case shows that the explanation of moral responsibility rests behind the eyes of the transgressive agent. Janus deserves to feel guilty. That offers you an instrumental reason to respond to him in whatever way happens to elicit this deserved guilt, which in this case is (oddly but by no means unimaginably) by turning the other cheek.

Blaming Janus can still be justified by other considerations. Blaming him may deter others. Perhaps retaining a proneness to blame in reaction to wrongdoing is, even when that blame is undeserved, part and parcel of maintaining good interpersonal relationships amongst most members of the moral community.[265] You may also justifiably be angry *that* Janus has done wrong in the way that we are justifiably angry that there is injustice in the world. That may give you a justification to voice your anger

---

[265] This is one way to read Strawson 1982.

to him, in a way blaming his action rather than him.[266] You may even have a right to blame him, given agreed-upon social norms.

These justifications to blame are consistent with the conclusion that Janus himself does not *deserve* to be blamed, and furthermore that allegedly deserved blame is merely a means to elicit deserved guilt. I suspect that the prevalence of the blame-focused view is due in part to a conflation of these subtly distinct justifications to blame.[267] Our (often overwhelming) reasons to express blame to an agent lead us to believe that it is our blame that the agent deserves. But an agent's desert is theirs and theirs alone. That the guilt a transgressive agent deserves is often the result of meting out independently justified blame is a contingent fact.

3 Summary and some objections

The aforementioned cases show that moral responsibility is resilient. Blame-insensitive agents like Janus can be morally responsible. So can suberogators and those who don't understand second-personal reasons. Blameworthiness is too fragile to account for moral responsibility in these cases. Guiltworthiness is not. It can account for moral responsibility in these cases, as well as in all those cases in which morally responsible agents are allegedly blameworthy. So we should abandon the blame-focused view and instead endorse

---

[266] Gideon Rosen claims that wrongdoers like Janus are indeed blameworthy. Suggestively, the only support he offers is that the "*act* remains resentment-worthy" (2015, 69). That conclusion is consistent with the conclusion that Janus himself is not blameworthy.

[267] It is also due, as I explain in "The Case for Retributive Caution", to our retributive and unconsciously self-affirming nature.

The guilt-focused view: An agent is morally responsible in a liability sense for a transgression just in case s/he deserves to feel moral guilt for that transgression.

There are a few remaining ways that an advocate of the blame-focused view might resist. One is to argue that, although the characters cited above do not deserve to be targeted by blame, they deserve you to privately *feel* blame towards them. As I pointed out, private blame towards attributable agents is morally permissible, given that it only affects the blamer. It is also understandable that you'd feel angry in some of these cases (perhaps all of them if you're a particularly irritable sort). But privately blaming is, as the old saying goes, like drinking poison and expecting another to sicken. The silliness of such an act impugns the plausibility of the position recommending it. In privately blaming, you're just hurting yourself. Janus, Lisa, Robinson, etc don't deserve a response that only hurts you. If an agent does not deserve to be targeted with blame, then s/he doesn't deserve to be privately blamed either.[268]

Another argument for the blame-focused view in light of the preceding criticisms is to broaden it to include desert of other negative interpersonal responses. Blame, the blame-focused advocate might claim, is merely the paradigmatic interpersonal response that is deserved for wrongdoing. It is not the *only* kind of negative interpersonal response that can be deserved. In Janus's case, blame is not deserved, but an alternative interpersonal response - turning the other cheek - is deserved. And while this interpersonal response is unlike blame in that it does not alienate, it is quite like blame in that it is supposed to communicate legitimate moral demands, and elicit guilt and

[268] Watson 1996, 239.

repentance. Similarly, although Lisa does not deserve to be blamed, she deserves her mother's cool questioning.

This amendment does not generate the right answer in the Robinson case. There is no interpersonal response that plausibly explains Robinson's moral responsibility. But there is a deeper problem. We are attempting to discover commonalities between all deserved responses for immoral acts. On this proposal, one commonality is the interpersonal nature of these responses. Another, as I hope I've established, is the guilt that the interpersonal response must bring about. But now the proposal seems to be that: An agent is morally responsible in a liability sense for a transgression just in case s/he deserves both guilt, and whatever interpersonal response happens to elicit this guilt. The interpersonal feature of the response looks unimportant. Guilt is doing all the work.

What of blameworthiness? I have used the Janus and Robinson cases to urge that blame and other such interpersonal responses are never deserved *per se*. Rather, they are justified by an agent's desert only as a means to elicit the response that the agent deserves. The guilt-focused view is by design consistent with this sweeping conclusion. But it is also consistent with the possibility that blame and other such interpersonal responses can be deserved. I have only argued that, because not all morally responsible wrongdoers deserve interpersonal responses, interpersonal responses do not themselves *explain* moral responsibility. Whether these responses can ever be deserved I leave in this paper as an open question.[269]

---

[269] Though I argue in "The Case for Retributive Caution" that, in fact, we ought to be skeptical that blame and other such censures are ever deserved.

Some may fret over the fact that the guilt-focused view requires the abandonment of a Strawsonian explanation of moral responsibility. Strawsonians allege that (i) it is some fact about the value of moral-interpersonal relations of the moral community that explains any given agent's blameworthiness for a transgression, and that (ii) blameworthiness explains moral responsibility for a transgression. Strawsonians have adeptly argued for (i). But the jump from (i) to (ii) is rarely justified. I have relied on Strawsonian insights about the strictness of certain interpersonal norms to argue that agents like Lisa and Robinson are not blameworthy for their transgressions. If I am right that it is these agents' guiltworthiness that explains their moral responsibility, and it is *also* true that the value of moral-interpersonal relations cannot account for their guiltworthiness, then the jump from (i) to (ii) is indeed unwarranted. Strawsonians cannot in that case account for moral responsibility.

However, Strawsonians still have an important theory even if it turns out that they cannot explain moral responsibility for transgressions. We need a theory that tells us how to respond to any given morally responsible transgressor. Strawsonians have offered influential arguments that we sometimes have strong reasons to proportionately blame or even punish[270] any given morally responsible transgressor. (To that list I urge Strawsonians to also add strong reasons to 'guilt' guiltworthy agents in a non-blaming, non-punitive manner.) And they have adeptly identified the norms governing these reasons. These reasons and the norms governing them obtain even though Strawsonians cannot explain moral responsibility.[271] So though it turns out that

---

[270] Feinberg 1965; Primoratz 1989; Kleinig 1991.

[271] McKenna 2013, 154-161; Pereboom 2014, 137-138.

Strawsonians can't explain moral responsibility, they still have a compelling theory that tells us something at least as important, if not more so; how to respond to the morally responsible. Perhaps that is no consolation.

Some may object that my arguments do not count in favor of the guilt-focused view so much as they count against the claim that moral responsibility as liability is explained by desert. They may object as follows. We ought not begin with the assumption that moral responsibility for transgressions is explained by desert, and then attempt to discover the response that desert for transgressions always justifies. Rather, we ought to begin with the assumption that moral responsibility for transgressions is explained by defeasibly justified blame, and then attempt to discover the defeasible justification at issue. The arguments in favor of the guilt-focused view show that desert for transgressions does not always offer a defeasible justification to blame. So desert can't be the justification at issue.

On this view, if one of the numerous non-desert-based justifications to blame an agent obtains, the agent is blameworthy. For me to deny blameworthiness only because of the absence of a desert-based justification splits hairs too finely. This objection fails for two reasons. First, one desideratum of any good theory of moral responsibility is an explanation of why moral responsibility for transgressions quite often offers a defeasible justification for blaming the transgressor. The objection clearly satisfies this desideratum. But another desideratum is an explanation for why our attributions of moral responsibility are typically couched in desert language, and bring along with them strong intuitions about desert. We should strive for a view that satisfies both desiderata. The guilt-focused view does just this. It takes our intuitions and linguistic practices to be

correct; we say and judge that the morally responsible deserve because the morally responsible deserve. Moreover, the guilt-focused view allows that moral responsibility for transgressions quite often (perhaps nearly always) offers a defeasible justification for blaming the transgressor. To retain the assumption that moral responsibility for transgressions *always* offers a defeasible justification to blame comes at the expense of the correctness of our judgments and linguistic practices concerning desert. That is too high a cost.

Second, there is a kind of case that militates against a more inclusive conception of blameworthiness as explaining moral responsibility.

> *Stanley*: Stanley and Meredith have been married for fifty-seven years. Their love for each other is unconditional. So after Meredith's nearly immobilizing stroke, Stanley dutifully looks after her every need. Still, the task is onerous for Stanley. He's elderly. He and Meredith have no children, relatives, or friends to help him. So Stanley often gets impatient. During one of his more impatient moments, Stanley doesn't bother to buckle Meredith in to the passenger seat of their car. He knows the danger this poses to Meredith, especially considering that his car has been acting funny lately. He understands all of the reasons why failing to prevent this danger is wrong. But he's in too much of a hurry to care. Moreover, his patience has been worn so thin lately that he has come to desire, albeit weakly, that Meredith die so as to relieve him of his immense burden. During their commute, the brakes give out, and Stanley's desire is unfortunately fulfilled. Meredith dies instantly during the crash. Stanley's guilt over wronging Meredith is

> so immense and relentless that he is unable to sleep, unable to eat, unable to function, unable to carry on a day without tears. He wishes for all the world that he had done otherwise, and would eagerly trade places with Meredith if he could. This guilt renders him incapable of any other emotions, and remains with him for the short remainder of his life.[272]

For Robinson, blame is unintelligible. For Janus, blame is ineffective. For Stanley, it is extraneous in that it can't bring about any effect that has not already occurred. Specifically, blaming Stanley can't communicate any moral demand that he isn't already satisfying. Nor can it inflict any more of the typical harms of blame upon him. He already feels as guilty as it is possible to feel. It seems to me that, though it is apt to morally disapprove of Stanley, there is *no* justification to blame him. On even a more inclusive blame-focused view, that renders Stanley not morally responsible for his wrongdoing. However, Stanley is clearly morally responsible for his wrongdoing. Stanley deserves some amount of the guilt that he feels. And it is this guiltworthiness that explains his moral responsibility for freely, knowingly, and inexcusably doing wrong. Sometimes moral responsibility hides behind a blameless face.

---

[272] Thanks to Michelle Kosch for encouraging me to include this kind of case.

Retributive Skepticism and the Rationality of Anger

When we believe that an agent has acted immorally, we naturally tend to react with anger.[273] We are, of course, sensitive to the fact that angry reactions are almost always harmful to the target,[274] and therefore require justification. Commonly, the justification offered is retributive. The immoral agent is alleged to *deserve* it, where by this it is meant that there is a non-instrumental justification to target the agent with anger *just* because she acted immorally in the way that she did.[275] However, this putative justification is often lacking. Some are undeserving of anger because they were coerced, compelled, or morally unaware.[276] Others are simply too psychologically abnormal to be eligible for desert at all. Moreover, there are excellent reasons to doubt that anyone can deserve to be targeted with expressions of anger or other harmful treatment. After all, the epistemic origins of desert attributions are notoriously suspicious.[277] And it is unclear that beings like ourselves have the kind of free will[278] and moral sensitivity[279] requisite for desert. On the basis of these worries, many of us are

---

[273] Strawson 1982; Wallace 1994, 76-77 and 1996, 18-50; Watson 1996, 238; Bennett 2002 and 2008; Smith 2007, 477; Strawson, G. 2008, 29-30; Nichols and Prinz 2010, 111-146; McKenna 2012, 21-29, 64-66; Boxer 2013; Rosen 2015.

[274] Feinberg 1984, 37; Bennett 2002, 149-153; McKenna 2012, 136-141; Boxer 2013, 125.

[275] Feinberg 1970, 55-94; Strawson, G. 2002, 452; Clarke 2005, 21; Pereboom 2009, 170 and 2014, 2; Waller 2011, 2-5; McKenna 2012, 150; Scanlon 2013, 101- 102.

[276] Strawson 2008.

[277] Nietzsche 1998; Murphy 2007, 17; Waller 2015, 39-52.

[278] Pereboom 2001; Honderich 1988; Strawson, G. 2008; Parfit 2011, 259-272; Strawson 1994; Rawls 1991, 273-277; 2003, 72-80; Scanlon 1988, 197-201; Scheffler 1992; Shafer-Landau 2000; Boonin 2008; Murphy 2007.

[279] Levy 2009; Rosen 2004.

becoming increasingly doubtful about desert as a justification for anger. What implications, if any, do retributive doubts have for the rationality of our anger?

That depends. There are many who claim that it is incredibly difficult, perhaps impossible, for an agent to target an individual with anger without also believing that the individual deserves it.[280] I call this view *retributivism about anger* or simply *anger-retributivism.* If this view is correct, then any agent who targets an individual with anger while skeptical about that individual's desert will be doxastically irrational in that she holds a belief that she believes to be false.[281] This is troubling. For a skeptic who wishes to avoid irrationality would then face a dilemma: either relinquish doubt or eschew anger. Some argue that the first option is best in light of anger's desirable consequences.[282] Unfortunately, this will often leave us forming beliefs in the teeth of contrary evidence. That being the case, others advise eschewing anger, opting instead for "feeling disappointed, hurt or shocked about what the offender has done, moral concern for him, and moral sadness and sorrow".[283] But this option requires a rather drastic emotional change, one that many consider to be undesirable or even impossible.[284] What's a skeptic to do?

In this paper, I argue that, although anger-retributivists have rightly identified the retributive nature of paradigm forms of anger such as resentment and indignation, they are mistaken that anger in general has any unbreakable connection to retributive

---

[280] Honderich 1988, 583-585; Smilansky 2002, 501; Scheffler 2003; Strawson, G. 2008; Pereboom 2014, 128-129.

[281] Pereboom 2001, 97 ftn.17.

[282] Vargas 2013.

[283] Pereboom 2014, 146.

[284] Wallace 1996; Strawson 2008; Nichols 2007; McKenna 2012; Boxer 2013.

beliefs. I first more carefully explain the anger-retributivist thesis, as well as the phenomena offered to support it. I then defend an alternative account that explains these phenomena just as well, but lacks the troubling implication. I call it the *practical view of anger*. Roughly, this is the view that a moral agent can target an individual with anger if she believes she is sufficiently justified in doing so, and values acting on that justification more than she disvalues harming the targeted individual. Due to the contingent details of our moral development, desert is the justification most often assumed to obtain, as well as one we highly value acting on. Hence the typical connection between anger and retributive beliefs. However, I argue that there are non-desert based justifications and values that can work just as well. I cite as paradigms two kinds of cases: cases of moral reform in which we value the betterment of the immoral agent and moral community more than we disvalue the temporary harm that anger causes, and cases of moral protest in which we react angrily to an immoral but nevertheless undeserving actor in order to overcome or take a stand against the immoral system of which they are a part. Given our values, lacking the belief that such individuals deserve to be targeted by anger might decrease the severity of our anger, but it surely cannot extinguish it. I conclude by discussing the upshot for retributive skepticism.

1 Desert-based and desert-free anger

First, some distinctions. I'll classify a belief as 'involved' with an emotion whenever it is a constituent, a cause, or a concomitant of that emotion.[285] Some think

[285] Prinz 2004, 3-20.

that anger needn't be involved with any belief at all.[286] I'm sympathetic to the view. But I'll be setting it aside, as I'll only be dealing with a kind of anger known as 'reactive anger'.[287] This kind of anger (henceforth simply 'anger') is necessarily involved with the belief that the targeted agent has acted immorally. As stated at the outset, a popular justification for targeting an agent with anger is that she deserves it. Call anger that is involved with the belief that the target of anger deserves to be targeted by anger *desert-based anger*. Call anger that lacks involvement with such a belief *desert-free anger*.

I grant for the sake of argument that paradigm forms of anger such as resentment and indignation are desert-based. According to Joel Feinberg, these forms of anger "are not mere 'reactions to' but 'requitals for'. [They] have ostensible desert logically built in to them", and are *not* targeted at agents "'in the public interest' or 'for utilitarian reasons'."[288] Derk Pereboom concurs, claiming that resentment and indignation are kinds of "anger targeted at an agent because of what he's done or failed to do," and necessarily come along with "a belief that the agent deserves to be the target of that very anger just because of what he has done or failed to do. An attitude does not count as resentment or indignation if it lacks these features".[289]

Some accept that anger can be desert-free, and claim that it is only certain forms of anger that are desert-based. Michael McKenna, for example, claims that it is only "reactive anger of a distinctive sort, specifically, resentment, indignation, disapprobation,

---

[286] Gibbard 1992.

[287] See especially McKenna 2012.

[288] Feinberg 1970, 70-71.

[289] Pereboom 2014, 128. Also see McKenna 2012, 66.

and guilt, that carries with it a basic-desert belief."[290] For others, there are no viable alternatives to desert-based anger. For example, Galen Strawson argues that all instances of anger are "demonstrably inappropriate given their essential dependence on a belief in [desert] that is demonstrably false".[291] Saul Smilansky claims that retributive skepticism "does not leave sufficient moral and psychological 'space' for...defensible reactive attitudes".[292] Samuel Scheffler claims that "the reactive attitudes entail...the thesis that those very attitudes are merited or deserved as responses to the individual who is their target".[293] Ted Honderich claims that there is a non-cognitive but nevertheless unavoidable connection between anger and retributive beliefs; due to human psychology, it is just a "brute fact" that retributive beliefs are concomitants of anger.[294]

One way to interpret these remarks is that desert-free anger is impossible *simpliciter*. But that's clearly too strong. You're walking along the sidewalk in heavy rain. It's been a long and frustrating day. A driver hits a nearby puddle trying to park, drenching you with foul water. You immediately have an increase in heart rate and skin temperature, distinctive facial expressions, and hostile action tendencies towards the driver[295], all of which are involved with your belief that the driver has acted immorally.

---

[290] McKenna 2012, 66.

[291] Strawson, G. 2008, 90. Strawson's reference in this passage is to 'true responsibility', which he has elsewhere clarified as desert-entailing. It is "responsibility and desert of such a kind that...can exist if and only if punishment and reward can be fair or just without having any pragmatic justification, or indeed any justification that appeals to the notion of distributive justice" (2002, 452).

[292] Smilansky 2002, 501.

[293] Scheffler 2003, 69-92.

[294] Honderich 1988, 583-585. Smilansky 2002, 501. Strawson, G. 2008, 90 Scheffler 2003, 69-92.

[295] Vidmar 2001, 43; Darley and Pittman 2003, 333-334; Nichols and Prinz 2010, 124.

Those conditions are sufficient for anger. But in that first moment, it's unlikely you have yet formed any beliefs about the deservingness of the target. That is because the anger is still "narrow-profile". *Narrow-profile* anger occurs instantaneously with the belief that one is being wronged.[296] It is a kind of anger that is not exposed to the whole of an emoter's cognitive system. We share narrow-profile anger with some of those who are cognitively impaired and children too young to be plausibly attributed with beliefs about desert.

Most of us become more reflective after the initial surge of anger though. That is because, for normal adults, narrow-profile anger soon becomes *wide-profile* in that it is exposed to a broader range of the emoter's attitudes, and is amenable to rational assessment.[297] Once an emotion is wide-profile, a person can appraise it as irrational, imprudent, or unjustified. Alternatively, one can appraise it as rational, prudent, or justified. It is at this appraisal stage that a belief about desert may become involved with one's anger.

Still, even at the wide-profile stage, it is implausible that a retributive belief is always involved with anger that does not result in action. Suppose after the initial surge of anger you discover that the driver was too distracted by his noisy family to have seen the puddle. The driver was in the wrong, but unknowingly. Consequently, you come to doubt that the driver deserves to be targeted with anger. That being the case, you're unlikely to act in a way that is hostile towards the driver. More specifically, your hostile action tendencies are in this case insufficient for action in the absence of a retributive

---

[296] Nichols 2007, 413-414.

[297] Nichols 2007, 412-416.

belief. But those action tendencies remain nonetheless, as well as the phenomenological and physiological features of anger, and the belief that the driver has acted immorally. You're silently angry. Your silent anger is desert-free. The above remarks by Strawson, Pereboom, and others are, then, more plausible when applied only to wide-profile anger at an individual that results in a hostile action towards that individual, or (as I'll stipulate is equivalent) results in targeting the individual with an expression of anger.

Some of the above remarks indicate that desert-free wide-profile anger is *impossible* to act on. That's also too strong. If desert-free wide-profile anger is in some way problematic, it is surely not because there is no possible world in which an agent can act on it. A more plausible thesis is that such anger is too difficult for normal adults to act on to be at all practical. Consider higher-order thoughts. We sometimes think about our thoughts. And occasionally these second-order thoughts are the object of some further third-order thought. A 20th-order thought is possible, but not the sort of thing that humans can do without a great deal of difficulty. Facts about human psychology render them psychologically impractical.

In response to these considerations, Derk Pereboom concedes that narrow-profile anger "has no cognitive content or presupposition or associated belief that involves the notion of desert".[298] He is also uncommitted about the possibility of children and cognitively unsophisticated adults expressing desert-free anger. But he contends that, at the very least, "when a mature, normal human being makes some agent the

---

[298] Pereboom 2014, 147.

target of an overt expression of her genuine [wide-profile anger[299]], it's at least close to psychologically impossible that she doesn't also believe that the agent basically deserves [it]".[300] In light of these suggested amendments, I take the most plausible interpretation of the above remarks to be what I call *retributivism about anger* or

> *Anger-retributivism*: it is not psychologically practical for a normal adult to target an individual with an expression of desert-free anger that is wide-profile.

This thesis still entails the troubling conclusion that normal adults will be doxastically irrational to doubt that an individual deserves to be targeted with anger while simultaneously targeting that individual with an expression of anger that is wide-profile. For if anger-retributivism is true, that very anger comes along with the belief that the target *does* deserve to be targeted with anger. This inconsistency between doubt and belief indeed generates a strong reason for the skeptic to either relinquish doubt or eschew anger.

2 Explaining the Harris Case

The primary support offered for anger-retributivism concerns the phenomenon of anger diminishment after the retraction of retributive beliefs. Consider the much-discussed case of Robert Harris. In 1978, twenty-five year-old Robert Harris decided, rather capriciously, to murder two teenage boys. Mere minutes after murdering the two

---

[299] In conversation.

[300] Pereboom 2014, 129.

teens, Robert Harris nonchalantly ate one of their hamburgers, and seemed to be "in an almost lighthearted mood. He smiled and told [his brother] Daniel that it would be amusing if the two of them were to pose as police officers and inform the parents that their sons were killed".[301] Those of us who are morally attuned become angry at Harris for what he has done, and believe that he deserves (at the very least) to be targeted by an expression of anger for it. Now consider Robert Harris's circumstances.

> He was the most beautiful of all my mother's children; he was an angel", [his sister] said. "He would just break your heart. He wanted love so bad he would beg for any kind of physical contact. He'd come up to my mother and just try to rub his little hands on her leg or her arm. He just never got touched at all. She'd just push him away or kick him. One time she bloodied his nose when he was trying to get close to her"...Robert Harris's father was an alcoholic who was twice convicted of sexually molesting his daughters. He frequently beat his children ... and often caused serious injury. Their mother also became an alcoholic and was arrested several times, once for bank robbery...Harris had a learning disability and a speech problem, but there was no money for therapy...Harris was raped several times, his sister said, and he slashed his wrists twice in suicide attempts...Everyone in the family knew that he needed psychiatric help.[302]

---

[301] Watson 1987, 269.

[302] Watson 1987, 272-274.

As Watson, Pereboom,[303] Nichols,[304] and others[305] point out, the force of our anger diminishes after considering these further details. What explains the dissipation of the emotion? Pereboom claims that after learning of the causal antecedents of Harris's immoral action, we abandon the belief that Harris deserves to be targeted by an expression of anger for what he has done. He claims that "the best explanation for [the dissipation of anger] is that your retributive attitude presupposed the belief that the killer deserved, in the basic sense, to be the object of this attitude, and because you no longer have this belief, the attitude is deprived of the presupposition that sustained it".[306]

However, the absence of a retributive belief is not the best explanation for the dissipation of anger. Reconsider the silent anger case cited above. You refrain from acting with hostility towards the driver because you abandon your belief that the driver deserves to be targeted with anger. It's plausible that you can, and probably would, remain silently angry nonetheless. Harris's actions were far worse than the driver's. But we don't remain silently angry. We lose anger altogether. The absence of a retributive belief cannot explain this difference between the cases.

Shaun Nichols argues that the vivid details of Harris's horrific childhood evokes sympathy, and that it is difficult to remain angry at Harris when one feels sympathy for him.[307] So even if it is granted that we come to abandon our belief that Robert Harris deserves to be the target of anger, we do not have to appeal to the doubt in order to

---

[303] Pereboom 2007, 202.

[304] Nichols 2007, 411.

[305] Kane 1996, 84; Honderich 1988, 434-435.

[306] Pereboom 2007, 202. Also see Kane 1996, 84.

[307] Nichols 2007, 413-6. Also see Arpaly 2006, 31.

explain the dissipation of the emotion; our conflicting emotion of sympathy is alone sufficient.

The anger-retributivist might respond that the conflicting emotion alleged to diffuse anger is not unlikely to be generated by a re-appraisal of the case, where this reappraisal at least partially consists in one's abandonment of the belief that Harris deserves to be the object of anger. Watson points out in his discussion that we empathize with Harris by coming to see him as morally unlucky, and sympathize with him by coming to see him as a victim.[308] Nichols concurs with the latter point.[309] However, both Watson and Nichols go on to claim, correctly I think, that these reappraisals are each sufficient on their own to generate the countervailing emotion. One needn't first abandon one's retributive belief on the basis of one's appraisal of Harris-as-victim or Harris-as-morally-unlucky in order for sympathy to diminish one's anger. And it's not as if seeing a person as victim or seeing a person as morally unlucky necessarily precludes the belief that she deserves to be the target of anger. Nichols' explanation is the best explanation.

3. Motivating an alternative account

A common motivation for claiming that a belief is involved with an emotion is that it is required in order to explain the apparent inappropriateness of instances of that emotion. For example, one reason to think that fear is an emotion that is inappropriate when one is clearly not in danger is that fear is involved with the belief that one is in

---

[308] Watson 1987, 275-276.

[309] Nichols 2007, 415.

danger. Therefore, when one is clearly not in danger, and one is nevertheless afraid, one's fear is inappropriate, given that it is involved with a belief that is epistemically unjustified.[310] At first glance it would seem that the anger-retributivist is likely to have more traction with this sort of claim. As Galen Strawson points out, "it seems that they [the reactive emotions] stand in a sufficiently close relation to [retributive] beliefs...to depend for their correctness or appropriateness on the correctness of those beliefs".[311]

Targeting the undeserving with anger is typically inappropriate. However, there are cases in which targeting the undeserving with anger is not inappropriate. I explore such cases in detail below. For the moment, I wish to focus on the vast majority of cases in which anger at the undeserving *is* inappropriate, as I think those cases tell us something important about anger's relation to retributive beliefs. I take it to be quite difficult for any human who assesses themselves as a moral individual to commit an act that they know requires justification without also intuiting that they are justified in doing so. This thesis carries armchair plausibility as well as a good deal of empirical support. Decades of cognitive dissonance studies have found that those who assess themselves as moral individuals are so averse to acting in a way that conflicts with this assessment that *ad-hoc* justifications for their behavior will seem to them to obtain, often at the expense of deeply-held beliefs.[312] Given that most of us are moral beings who understand that acting on our anger requires justification, any expression of wide-profile anger challenges our view of ourselves as moral individuals in the absence of an

---

[310] See Kendal Walton's discussion of 'quasi-fear' (January 1978, 5-27).

[311] Strawson, G. 2008, 92-100. Also see Strawson, G. 1986, 80; Murphy 1988; Scheffler 1992, 313-316 and 2003, 71.

[312] See especially the review by Joshua Knobe and Brian Leiter (2007).

intuition that the expression of anger is justified. So, an expression of anger will typically be involved with an intuition that expressing anger is justified. We tend to believe what we intuit. So, an expression of anger will typically be involved with the belief that expressing anger is justified

In most cases in which we are angry, there is no non-desert based justification to target a wrongdoer with anger. That she deserves to be targeted 'just because' she did wrong is often the only putative justification available. This explains why anger not only seems to be, but often *is* inappropriate when a desert-based justification is clearly lacking; our anger is involved with the belief that the anger is justified, and in most cases we believe, correctly, that there are no desert-free justifications available. Hence the inappropriateness of the anger when we believe that a desert-based option is also unavailable. We've simply run out of justifications.

My claim, then, is that it is psychologically difficult, perhaps impossible, for a normal adult who assesses herself as moral to target an individual with an expression of wide-profile anger without also believing that she is justified in doing so. This explains why anger at the undeserving is often inappropriate. What about cases in which an emoter rightly believes that they have some desert-free justification available? Here I wish to concede that the belief that one has a desert-free justification to express anger is unlikely to offer sufficient motivational force against desert-based values. Most of us disvalue harming an agent if we believe that she does not deserve it. One reason for our natural reluctance to accept consequentialist theories of punishment is precisely because this disvalue tends to generate an emotional aversion to the conclusion that it is sometimes obligatory to harm those apparently undeserving of it. This is especially

true when the consequences of the harm are only marginally better than they would have been in its absence. Without values of sufficient strength, our emotional aversion to treating persons as mere means will often keep many of us from actually targeting the non-deserving with anger even when we are clearly justified in doing so. That is why I endorse what I call *the practical view of anger,* or

> *Practicalism*: it is psychologically practical for a normal adult to target an individual with an expression of wide-profile anger if (i) she believes she is sufficiently justified in doing so, and (ii) values acting on that justification more than she disvalues harming the targeted individual.

3.1 Protest anger

Take an interaction Frederick Douglass describes in his autobiography. He writes of the aftermath of his first fight against a brutally oppressive slave-master who attempted to whip him for no good reason. It was a fight that lasted "nearly two hours," and carried with it the risk of death for Douglass.

> The whole six months afterwards, that I spent with Mr. Covey, he never laid the weight of his finger upon me in anger...This battle with Mr. Covey was the turning-point in my career as a slave. It rekindled the few expiring embers of freedom, and revived within me a sense of my own manhood. It recalled the departed self-confidence, and inspired me again with a determination to be free. The gratification afforded by the triumph was a full compensation for whatever

> else might follow, even death itself. He only can understand the deep satisfaction which I experienced, who has himself repelled by force the bloody arm of slavery. I felt as I never felt before. It was a glorious resurrection, from the tomb of slavery, to the heaven of freedom. My long-crushed spirit rose, cowardice departed, bold defiance took its place; and I now resolved that, however long I might remain a slave in form, the day had passed forever when I could be a slave in fact. I did not hesitate to let it be known of me, that the white man who expected to succeed in whipping, must also succeed in killing me. From this time I was never again what might be called fairly whipped, though I remained a slave four years afterwards. I had several fights, but was never whipped.[313]

Douglass speaks of deservingness occasionally in his autobiography. And it is unlikely that he lacked the belief that Covey deserved to be targeted by anger. But notice that Douglass never explicitly attributes his oppressor with deserving anything at all, much less as deserving to be targeted by anger or any other detrimental treatment. Rather, Douglass vividly recounts only the terribleness of Covey's oppression, and the good consequences that came about from angrily fighting against him in response to that oppression. His justification at the time of the fight was that Covey "had used me like a brute for six months, and that I was determined to be used so no longer"[314]. The former conjunct indicates the immoral action to which Douglass is responding. The latter contains a response that is instrumental to stopping any further moral transgressions.

---

[313] Douglass 2009, 78. For a discussion of the merits of Douglass's anger, see Bell 2009.

[314] Douglass 2009, 77.

Though it's Covey in specific that Douglass is reacting to, it's fighting off "the bloody arm of slavery" and a determination to not be used "like a brute" that is his justification for the reaction. Douglass's overt expression of anger was not mere narrow-profile anger. It was sustained. It survived rational reflection. It was wide-profile anger.

I take it that someone in a position like Douglass's could target a person like Covey with an overt expression of this anger while also lacking the belief that the target deserves to be the target of the anger. After all, the valuable consequences of this overt expression for Douglass were obvious and overwhelming. And even if someone in Douglass's position, perhaps Douglass himself, would have come to lack the belief that his oppressor deserved any sort of harmful treatment, it is almost certain to be the case that he would find it difficult, perhaps psychologically impossible given his situation, to value Covey's well-being enough for this disbelief to be motivationally salient. For a person in Douglass's position, the instrumental value of making a vicious oppressor like Covey worse-off far outweighs the value of Covey's well-being. This weighting of values surely has strong motivational force.

With this in mind, it seems psychologically practical, even easy, for a person in such a situation to coherently reason that, "I'm angry at Covey for his lifetime of viciousness towards me and numerous others. Given the conditioning of his upbringing and social environment, I don't believe that Covey himself deserves to be targeted by this anger. But I don't care much about Covey's well-being, or of his lack of deservingness. Nor should I. Instead, I deeply value the end of Covey's oppression. And I am certainly justified in making Covey worse-off by acting on my anger as a means to these far better ends." Again, I'm not claiming Douglass himself had these beliefs, or

reasoned or emoted in this way, but merely that these circumstances could easily elicit such reasoning and emoting.

An anger-retributivist might allege that someone in Douglass's position would have the relevant retributive belief, but simply fail to realize it. I don't find this response compelling. Although there are cases in which beliefs are not introspectively available, it is implausible that a retributive belief would elude an emoter in such a case. After all, if such an emoter held a belief that she believed to be false, she would surely feel *some* psychological tension. But it is intuitive that a retributive skeptic in Douglass's position would feel no psychological tension targeting a vicious oppressor like Covey with an expression of desert-free anger. So there is not good evidence that the emoter holds a retributive belief. In the absence of such evidence, it would be uncharitable to attribute such a belief to the emoter, as doing so would amount to attributing a seemingly rational person with irrationality.

The kind of scenario envisaged here is not rare. Many wrongdoers, regardless of circumstances undermining their desert, are part of oppressive systems too terrible to elicit any sympathy. They are part of social systems that wantonly harm by way of fatuous prejudice, ruthless psychological abuse, or violent oppression. They are racists, bigots, domestic abusers, bullies, and despots. The systems of which such individuals participate, and the individuals themselves, don't always respond to sorrow or pleas for mercy. Sometimes, anger is a much more effective tool.[315]

3.2 Reformative anger

---

[315] Bell 2009; Nichols 2007.

Consider next anger that is reformative. There are many cases in which we believe that, despite some amount of immediate harm, an agent will be made much better-off overall by being targeted by an overt expression of anger. Consider the following case.

> Because of the turmoil of a recent divorce, Jim's teenage daughter Sandra begins stealing from the corner store. This angers Jim. Jim has refrained from overtly expressing his anger because he rightly believes that doing so will hurt Sandra's feelings, and that Sandra, being compelled to steal by the turmoil at home, does not deserve any adverse treatment. Instead, he disciplines her as best he can in other ways, expressing sorrow and remorse all the while. But she is not particularly receptive. She continues to steal. Jim comes to believe that hurting Sandra's feelings by way of an overt expression of his anger will compel her to refrain from theft, and that this will make her much better off in the long run. Jim believes this justifies expressing his anger and, being a loving parent, values his daughter's being made better-off overall far more than he disvalues the small amount of her displeasure that is a means to that end.

My intuition is that Jim can express desert-free anger. There is a concern that this kind of anger is merely 'feigned'. But there are crucial physiological and psychological differences between feigned anger and reformative anger. A paradigm example of feigned anger is the outburst of a professional stage actor. A stage actor's feigned anger may be behaviorally identical to genuine anger. And yet an actor's adrenaline, heart

rate, and skin-conductance are unlikely to increase, and there won't be any motivation to harm the target of the performance. An actor would be shocked if her outburst were to actually hurt her fellow performer. In contrast, every time Jim catches Sandra stealing, he develops the physiological and psychological symptoms that feigned anger lacks. He's not acting like he's angry. He's actually angry. This happens to parents all the time when their children act wrongly. Jim may not enjoy that Sandra is hurt by his expression of anger, but he expects that it will, and would be shocked if it didn't.

I grant that, phenomenologically, there is likely to be a difference between anger aimed at an agent for reformative purposes, and anger aimed at an agent as retribution. Specifically, I think the expression of retributive anger tends to feel good, while anger expressed solely for reformative purposes probably does not. This is because only the latter is aimed at helping the target, and as such is involved with personal values that speak against the momentary harms that are inflicted. Jim, for example, is acting on his anger as a last resort for precisely this reason. Although this is an interesting difference between reformative anger and retributive anger, it is not one that is relevant to the debate.

It's an open question as to just how often anger would be helpful in this way. Christopher Bennett, Michael McKenna, and Karin Boxer have all recently argued that the harm of being targeted with anger tends to bring about much better circumstances overall for the immoral agent, as well as the members of her moral community. It can bring about a sort of healing process between the immoral agent and her victim involving remorse, apology, and ultimately reconciliation.[316] I am highly skeptical that

---

[316] Bennett 2002; 2008; McKenna 2012; Boxer 2013.

anger is typically more conducive to such effects than non-hostile alternatives. However, given that it is common amongst many cultures to seek apology and reconciliation by way of overt expressions of anger, and that most immoral agents are reverent to these cultural norms, it seems plausible (though this is a speculation that can only be corroborated by empirical research) that in a small but substantial number of cases an expression of anger will result in more good than any non-hostile alternative.

4. Conclusion

The anger-retributivist may not share my intuitions, and may object that all these examples do is pit practicalist intuitions against theirs. However, I take it as a source of agreement that we ought to strive for parsimony in theorizing about psychological states. Practicalism attributes fewer beliefs to emoters than anger-retributivism, but attempts to explain the phenomena under investigation just as well. It is the simpler view. In an intuition standoff, parsimony puts the onus on anger-retributivists to provide something aside from their particular intuitions to defend their more complex view. My hope is that I have argued persuasively in section 2 against the support that has been offered by anger-retributivists thus far. If so, the ball is now in their court.

The results of anger are typically bad. Among other things, it can distort our beliefs. It can damage personal relationships. It can lead to childish and destructive behavior. So it is often the case that anger lacks any desert-free justification. Retributive skepticism undermines the only remaining putative justification. That being so, abstaining from anger will often be the best course of action. However, not all expressions of anger lack justification in the absence of desert. We value acting on

some of those justifications - for example those in the interest of moral protest and reformation - far more than we disvalue the harm caused to the target of anger. Such anger is not only justified. It is practical and rational. In these cases, desert-free anger is for the retributive skeptic like lactose-free milk is for the allergic; perhaps it's not as tasty, but it serves just as well.

Appendix: In Defense of *BRW*

As stated above, I think it's plausible that

> *BRW*: an agent deserves to be blamed for ϕ-ing only if it is wrong that s/he ϕ's all-things-considered.[317]

Two different kinds of arguments have been offered against *BRW*. The first relies on the 'subjective view' of blameworthiness. The second relies on a conclusion alleged to be supported by Frankfurt cases. I explain and argue against each in turn.

1. The subjective view of blameworthiness

Consider the following case, offered by Ishtayique Haji.

> *Cure*: Suppose doctor Deadly is responsible for the treatment of some patient, Bennie. Suppose all the available evidence indicates that Bennie is suffering from a dangerous disease that we will call 'Malady.' Suppose Malady can easily be cured by administering one dose of medicine A, but exacerbated to the point where it proves fatal if a patient suffering from it is given a single dose of medicine B. Suppose wicked Deadly gives Bennie medicine B with the express intention of killing Bennie. But luckily, suppose the diagnosis is incorrect. Bennie is in fact suffering from Malaise, a disease that gives rise to symptoms almost

[317] Widerker 1991, 223; Wallace 1994, 12, 135; Copp 1997, 445; Pereboom 2009, 170; Nelkin 2011, 100-107; Boxer 2013, 64. I intend ϕ-ing to include refraining.

> indistinguishable from the ones to which Malady gives rise. Suppose, finally, that a dose of B given to a patient suffering from Malaise safely cures Bennie, but a dose of A given to a patient stricken with Malaise instantly kills him.[318]

As Haji points out, Deadly's administering the cure is not impermissible. On the contrary, it is obligatory; Deadly *ought* to administer the cure, all things considered. Nevertheless, Haji claims that Deadly deserves to be blamed for administering the cure.[319] According to Haji, echoing a view by Zimmerman, the case supports the 'subjective' view of blameworthiness that desert of blame for an action does *not* require that an agent transgresses a moral obligation, but only requires that she "performs the action in light of the belief that she is doing wrong."[320] Hence, *BRW* is false.

I find this troubling. First, the subjective view entails false conclusions to other cases. A parent in the deep south may firmly believe, justifiably, given her upbringing and social circumstances, that she is obligated to brutally whip her son for his homosexual tendencies. Despite this firmly held belief, she may, due to purely

---

[318] Haji 1998, 51. Also see his 2002, 172, as well as Zimmerman 1997, 236; Scanlon 2008, 124-125.

[319] Haji 1998, 51.

[320] Haji 1998, 52. Also see Zimmerman 1988, 40-40. In the work in which he presents the above case, Haji endorses only the desert of the sort of aretaic blame - what I call 'moral disapproval' - I discussed at the outset. He says that "the blame in consideration is manifestly inward and not overt. It is the sort of blame registered when we say things like 'she is deserving of blame for pilfering the pie'; or 'she is to blame for letting the candle go out.' The fact that a person can be deserving of moral blame even in the absence of anyone else's being aware that she is so deserving underscores the inward nature of such blame" (Haji 1998, 9). However, in a more recent work, Haji speculates that "perhaps the praise and blame in question comprise overt praise and blame respectively. Such praise or blame consists in reaction to, or treatment of, a person on account of some deed performed by that person, the reaction being positive (like a pat on the back) in the case of praise, and negative (like a scowl or a reprimand) in the case of blame." He is then explicit that overt blame for an act does not imply that the deserving agent ought not to have performed it (Haji 2002, 40). Zimmerman certainly seems to have such blame in mind when he discusses liability for acts that are not wrong (1988, 180).

prudential considerations (such as squeamishness) not whip her son. Just as in Cure, the subjective view of blameworthiness implies that, since this parent acts in light of the belief that she is doing wrong, she deserves blame for not whipping her son.[321] This is an incredibly counterintuitive conclusion considering that this parent is morally forbidden from whipping her son. The correct conclusion is that this parent does *not* deserve blame for not whipping her son. *BRW* entails this conclusion.

Second, the subjective view of blameworthiness leads to an unfair sort of double-standard. Doctor Deadly 'ought' to have administered medicine B, all things considered. It was easy for Deadly to administer B, he had no moral reservations about doing so, and demanding him to do so would not have harmed him in any way. This being the case, it seems at least in principle fair to Deadly to demand him to administer B. If we accept Haji's conclusion that Doctor Deadly is worthy of blame for administering B, then it would seem that Doctor Deadly deserves to be blamed for an action that it was fair to demand that he do. But that is *unfair*. To demand Deadly to administer the cure indicates that the demander will not accept noncompliance with the demand. But to then blame Deadly for complying with that demand is to do just the opposite; it is to *not* accept compliance with the demand. That is needlessly unfair. Deadly would seem to be damned if he does and damned if he doesn't.

For these reasons and others, I have a strong intuition that Deadly does *not* deserve blame for administering the cure.[322] As David Copp argues, if a person like Deadly deserves blame, it is instead for being "willing to do something that he believed

---

[321] Haji 1998,

[322] One of the alternative desert bases sometimes cited in cases such as this is that Deadly deserves blame for *trying* to kill Bennie. Justine Capes (2012) has persuasively argued against that proposal.

to be wrong".[323] Dana Nelkin and Gideon Rosen further specify that a person like Deadly will be blameworthy when it is a terrible character trait or intention - for example maliciousness - that motivates his willingness to do something that he believes to be wrong.[324] Indeed, in discussion of a similar case, Justin Capes argues that an agent is blameworthy for committing a morally permissible action if she does it out of ill will, and "despite believing that it was wrong".[325] It is quite plausible on nearly any moral theory that being willing to maliciously do something one believes to be wrong is itself wrong.[326] That's the best explanation for blameworthiness in the sorts of cases that those endorsing the subjective view appeal to. But that explanation doesn't support the subjective view. So the subjective view of blameworthiness is unmotivated.

2. An appeal to Frankfurt cases against BRW

Haji offers a different sort of argument to the conclusion that blame can be deserved for bad actions and characteristics that are not wrong. It is as follows. An agent can transgress a moral obligation to not ϕ only if the agent ought not ϕ, all things considered. Following Kant's 'ought-implies-can' dictum[327], Haji argues, along with many others, that it is wrong for an agent to ϕ only if that agent can avoid ϕ-ing. But the characters in Frankfurt cases (to be discussed in detail below) could not avoid their deliberations, decisions, or actions. So characters in Frankfurt cases can't have done

---

[323] Copp 1997, 449.

[324] Nelkin 2011, 105; Rosen 2015, 76.

[325] Copp 1997, 448; Nelkin 2011, 105; Rosen 2015, 76.

[326] Copp 1997, 448;

[327] Kant 1998, 540-541.

wrong by deliberating, deciding, or acting as they did. However, it is widely agreed that the characters in Frankfurt cases can nevertheless deserve blame for deciding or acting as they did. So it follows that an agent can deserve blame for his action or characteristic even if that action or characteristic is not wrong.[328] More formally,

(1) It is wrong for an agent to ϕ only if that agent can avoid ϕ-ing.

(2) Agents in Frankfurt cases cannot avoid ϕ-ing.

(3) So (from 1 and 2), it is not wrong for agents in Frankfurt cases to ϕ.

(4) Nevertheless, agents in Frankfurt cases can deserve to be blamed for ϕ-ing.

(5) So (from 3 and 4) agents can deserve to be blamed for ϕ-ing even if it is not wrong for them to ϕ.

Hence, *BRW* is false. The argument is valid. But in the remainder of this section I argue that a careful investigation of Frankfurt cases shows that (4) is false. Two specifications are in order.

First, agents can be *derivatively* blameworthy for actions for which they had no alternative. A paradigm case concerns intoxication. Suppose one of Jones's patients has a heart attack. Jones is called in from a dinner party. If sober, Jones could easily fulfill his duty to save the patient's life. Unfortunately, Jones has been drinking. Because of his intoxication, he can't save his patient's life. Does Jones deserve to be blamed for failing to save his patient's life? That depends. At the time Jones decided to drink, he either had sufficient reason to believe his patient might need his help, or he didn't. If he

[328] Haji 1998, 53-59.

did, then he knowingly decided to deprive himself of the ability to save a patient that might need his help. In that case, it's intuitive that he deserves to be blamed for that decision. If Jones did not have sufficient reason to believe his patient might need him, then Jones does *not* deserve to be blamed for taking a drink. It seems Jones *derivatively* deserves blame for failing to save his patient's life only insofar as he *non-derivatively* deserves blame for this prior decision. Henceforth, the non-derivative desert of blame is what I will mean by 'blameworthiness'.

Second, the sense of 'can' that is relevant in this kind of argument was long thought to be nothing more than alternative possibilities *simpliciter*. That is why Harry Frankfurt originally argued against the 'principle of alternative possibilities'. A long debate about the issue revealed that the relevant sort of principle is more specifically the *principle of robust alternatives*

> *PRA*: *A* is blameworthy for ϕ-ing only if she had a *robust alternative* to ϕ-ing.

Derk Pereboom offers the following formulation of a robust alternative:

> *Robust alternative*: For an agent to have a robust alternative to an immoral action *A*, that is, an alternative relevant *per se* to explaining why s/he is blameworthy for performing *A*, it must be that
>
> (i) s/he instead could have voluntarily acted or refrained from acting as a result of which s/he would be blameless, and

(ii) for at least one such exempting acting or refraining, s/he was cognitively sensitive to the fact that s/he could so voluntarily act or refrain, and to the fact that if s/he voluntarily so acted or refrained s/he would then be, or would likely be, blameless.[329]

2.1 Pereboom's Frankfurt case

Several decades ago, Harry Frankfurt offered an intriguing style of thought experiment alleged to be a counterexample to the principle of alternative possibilities.[330] Derk Pereboom has ingeniously adapted Frankfurt's original thought experiment to both apply to *PRA*, and be insulated from many influential critiques. Suppose that

Grey: Adam is perched near the window of an abandoned warehouse, his rifle aimed at Bennie, the benevolent leader of a neighboring country. He's deliberating about whether or not to assassinate Bennie. His motivations to assassinate Bennie are strong. They're so strong that he will refrain from deciding to assassinate only if he becomes more attentive to a particular moral reason against assassination. Specifically, Adam will not decide to assassinate Bennie only if he becomes more attentive to the suffering the assassination will cause. Though Adam's attentiveness to this moral reason is necessary, it is not sufficient for him to refrain from deciding to assassinate Bennie. After all, he could still freely decide to assassinate Bennie, even while attending to this moral

[329] Pereboom 2014, 13. Pereboom considers decisions as kinds of actions, as do I.

[330] Frankfurt 1969.

> reason. Unbeknownst to Adam, a CIA operative, 'Grey' has been ordered to ensure that Adam's assassination succeed. Grey has the ability, by way of neurological manipulation, to control Adam's decisions. She intends that Adam decide to assassinate Bennie, and will cause Adam to do so if she sees that Adam becomes attentive to the moral reason against assassination. But as it turns out, Adam never attends to this moral reason. So Grey *never* has to intervene. Subsequently, Adam decides to act exactly as Grey had planned.[331]

Adam has the alternative available to him of becoming attentive to the moral reason against transgressing. But Pereboom argues that this alternative is not robust. Adam neither believes, nor has reason to believe, that becoming more attentive to the moral reason will trigger an intervention that will cause him to decide as he does. On the contrary, he deliberates and decides to act on the false assumption that he could refrain from deciding as he does. So the presence of Grey deprives Adam of any robust alternatives to his decision.[332] But many nevertheless intuit that agents such as Adam still non-derivatively deserve to be blamed for their decisions. Hence, *PRA* is false.

2.2 Ginet's critique

I remain unconvinced by such examples. Adam must decide by a certain time, say *t3*, or he will miss his opportunity to assassinate Bennie. So if Adam is going to

---

[331] Pereboom's original example involves tax evasion. However, considering the obviously immoral practices that tax dollars are used to support - drone strikes, for example - it is highly controversial that it is immoral to evade taxes. I appeal instead to an uncontroversially immoral action - assassinating a benevolent leader for no good reason. This 'Grey' case is in all other respects identical to the 'tax evasion' case presented in Pereboom 2014, 15. Also see Pereboom 2012.

[332] Pereboom 2014, 15-16.

refrain from deciding, he must become attentive to the moral reason at the time *t2* just prior to *t3*. So by *t3*, Adam has *already* either become attentive to the moral reason, or decided. If Adam has already become attentive, then Grey will have intervened by causing Adam to decide. In that case, Adam is *not* blameworthy for deciding as he does because his decision will have been caused by Grey. But if Adam has already voluntarily decided prior to *t3*, without any intervention, then it is clear that he is *already* blameworthy for deciding at the precise time that he does.

Carl Ginet argues, correctly I think, that this judgment does not conflict with *PRA*.[333] At the precise time that Adam decides - let us suppose it is *t1* - Adam could have instead been attending to the moral reason. This available alternative is robust at *t1*. For by voluntarily attending to the moral reason at *t1* Adam would have been refraining from deciding to do wrong. And Adam is surely cognitively sensitive to the fact that, if he refrains from deciding to do wrong at a certain time, he will not be blameworthy for deciding to do wrong at that precise time. So *PRA* is consistent with the intuition that Adam is blameworthy for deciding at the exact moment that he does.

On the other hand, many lack the intuition that Adam is blameworthy for deciding *by t3*. As Ginet points out, Adam is clearly blameless for the fact that, if he "had not [decided] earlier, he would have [decided] by *t3*."[334] That is instead a fact attributable to Grey. "But that fact is equivalent to the fact that [Adam decided] by *t3*."[335] Hence, Adam is not blameworthy "for the obtaining of the temporally less specific state of affairs of his

---

[333] Ginet 1996 and 2002.

[334] Ginet 2002, 308.

[335] Ginet 2002, 308.

[deciding] by *t3*".[336] Apparent intuitions to the contrary rest on a failure to carefully distinguish between being blameworthy for deciding at *t1* and being blameworthy for deciding by *t3*. This failure is to be expected considering that Adam would have been just as blameworthy, and for the same reasons, for the latter as he would the former, had Grey not been present.[337] Hence, *PRA* is not challenged by Frankfurt style examples.

2.3 Pereboom's response

Pereboom offers the following response. Ginet's timing criticism rests on the claim that any initial intuition that *PRA* is false is elicited by a failure to carefully distinguish Adam's deciding at *t1* and deciding by *t3*. According to Ginet, we fail to make this distinction because Adam would have been just as blameworthy, and for the same reasons, for the latter as he would the former, had a Frankfurt intervener not been in place. But consider the following case.

> *Bomb*: Adam knows he has the opportunity to set a bomb to explode exactly at *t4*, killing Bennie instantaneously. Suppose that factors beyond Adam's control causally determine him to decide to set this bomb, but leave it up to him at which instant during a short interval, beginning at *t0* up to and including *t3*, he makes the decision. More precisely, factors beyond his causal reach causally determine him to have a desire to kill Bennie so powerful that he will inevitably make this

---

[336] Ginet 1996, 407.

[337] Ginet 1996, 406-407.

> decision at some time in this interval. Adam first has this desire just before *t0*. It would persist to *t3* were he not to decide before then, and would not alter in strength during the interval. This last instant, *t3*, is the deadline because Adam believes, correctly, that the bomb is rigged not to explode if he decides any later. Due to Adam's being cognitively sensitive to the strength of his desire, from the time just before *t0* he believes with certainty that he will decide by *t3*. The bomb will explode at *t4* no matter which of these instants he makes the decision. Adam understands that at which of these instants he decides makes no difference morally, and as a result he is indifferent among them.[338]

Suppose Adam decides to set the bomb at *t1*. According to Pereboom, all incompatibilists are committed to the conclusion that Adam is blameless for deciding at this time. Because leeway incompatibilists like Ginet rely on *PRA* to show why determinism renders agents blameless, the reason that Ginet must cite for the conclusion that Adam is blameless for deciding at *t1* is that Adam lacks a robust alternative. However, Adam does *not* lack a robust alternative to deciding at *t1*. Rather, Adam has no alternative to deciding at *t3*, if he hasn't done so earlier. So "the only plausible candidate for explaining Adam's blamelessness is the unavailability of an alternative to making the decision by *t3*, and thus on the leeway incompatibilist view, this unavailability would have to be sufficient for Adam's not being blameworthy at *t1*."[339]

---

[338] Pereboom offers this case in correspondence as a cleaner version of his original 'Adam' case (2012 and 2014, 24-25).

[339] Pereboom 2014, 25.

Adam is not deprived of a robust alternative to deciding prior to *t3*. He is only deprived of a robust alternative to deciding by *t3*. So the leeway incompatibilist must also conclude that this lack renders Adam blameless at *t1*. Hence, in *Grey*, Ginet cannot consistently claim that Adam would have been just as blameworthy for deciding at *t1*, and for the same reasons, as he would be for deciding by *t3*, had a Frankfurt intervener not been in place. Still, the *Grey* case generates the "strong intuition that [Adam] is blameworthy at *t1*, for which the leeway incompatibilist now has no explanation."[340]

2.4 Against Pereboom's response

All leeway incompatibilists will agree that, in *Bomb*, Adam is not morally responsible for deciding by *t3* because

R: Adam has no alternative to deciding at *t3*, if he hasn't done so earlier.

However, leeway incompatibilists are not committed to the conclusion that *R* is the explanation of Adam's blamelessness for deciding to kill Bennie at *t1*. There's a vast difference between committing to ϕ, and deciding when to ϕ. Adam certainly commits to killing Bennie. But given the details of the case, that is a commitment he made at *t0*. Consider a case Pereboom offers as analogous to Adam's situation.

> Suppose my kids are very hungry, it's now *t0*, I need to leave at *t3*, and I am committed with certainty to feeding them before I leave. Suppose I can refrain

[340] Pereboom 2014, 26.

> from feeding them at *t1*. It's highly credibly that this is only because I know that I can feed them a little later, and not because I can refrain from feeding them at all by *t3*.[341]

The reason Pereboom can't refrain from feeding his kids by *t3* is because, given the strength of his attitudes and desires, and his sensitivity to them, he's already committed to feeding them by *t3*. For Pereboom, the question is no longer 'if' he will feed his kids. It is merely a question of 'when.' More generally, if an agent believes with good reason that her motivations for ϕ-ing will remain fixed in strength, and that those motivations guarantee that s/he will ϕ regardless of any further deliberation or efforts of her agency, then the agent has already committed to ϕ-ing.

An agent can be blameworthy for an immoral commitment if s/he formed it freely, but can't be if s/he didn't. To see why, reconsider *Bomb*. According to the setup of *Bomb*, Adam strongly desires to kill Bennie. He knows at *t0* that, because his current motivations for killing Bennie will remain fixed in strength, he will kill Bennie by *t3* regardless of any further deliberation or efforts of agency. For Adam, it is no longer a question of 'if' he will kill Bennie. It is merely a question of 'when'. So at *t0*, Adam is already committed to killing Bennie. Suppose Adam had a robust alternative to forming this commitment prior to *t0*. For example, suppose prior to *t0* he knew he could stop his desire to kill from becoming so strong by focusing on a moral reason against killing, but freely did not do so because of his own nefarious goals. In that case, it's intuitive that Adam can be blameworthy for forming the immoral commitment at *t0*. However, in

---

[341] Pereboom 2014, 25.

Bomb, Adam has no alternative at *t0* to forming the commitment to kill Bennie. So, given leeway sensibilities, Adam is blameless for committing at *t0* to kill Bennie.

Granted, Adam must now decide *when* to kill Bennie, or more precisely, when to act on his commitment to kill Bennie. But as Pereboom points out, because the bomb will explode exactly at *t4* no matter which time Adam chooses, it makes no moral difference when Adam decides to act on his immoral commitment to kill Bennie. So the fact that Adam decides to act on his immoral commitment at *t1* makes Adam no more or less blameworthy than if he had decided to act at some other time in the sequence. Adam is unavoidably committed to killing Bennie. That fact, and not *R*, renders him blameless. The strict deadline in *R* is, in fact, entirely irrelevant. Suppose that

> *Coma*: Adam is a hospital worker where Bennie will be in an induced coma for at least a year. Adam has a strong desire to kill Bennie starting at *t0*. Adam knows he's going to have sporadic opportunities to pull the plug on Bennie over the next year. He also knows that his strong desire to kill Bennie won't alter in strength. Adam has no desire to restrain himself indefinitely. That being the case, he knows he'll kill Bennie at some time in the near future regardless of any further deliberation or efforts of his agency.

On my view, Adam has already committed in *Coma* to killing Bennie at some time in the near future, even though he hasn't decided precisely when to act on this commitment, and there's no strict deadline for doing so. If Adam had some robust alternative to forming this commitment at *t0*, then it's intuitive that he can be blameworthy for forming

it. But if Adam has no robust alternative for forming this commitment at *t0*, it's intuitive that he's *not* blameworthy for it, despite there being no strict deadline. That's the conclusion entailed by *PRA*.

This 'unavoidable commitment' explanation is consistent with Ginet's timing criticism. Again, Ginet's explanation is that, in the *Grey* case, Adam decides at *t1* to assassinate Bennie with a robust alternative available. Intuitively, he's blameworthy for that decision. But it is not intuitive that Adam is blameworthy for deciding by *t3* to assassinate Bennie since that is a fact attributable to Grey. Notice that this explanation remains viable even if we stipulate that Adam forms a commitment to kill Bennie at *t1*, but takes a moment to better aim his gun and consequently does not act on his immoral commitment until after the *t2* deadline. Intuitively, he's still blameworthy. The reason is that, on this variation of the case, the basis of blameworthiness - his commitment to kill - has already occurred in the absence of any intervention, and in the presence of a robust alternative. *When* he decides to act on the commitment needn't matter.

In response to this criticism, Pereboom offers a reformulation of the Bomb case.[342] Suppose that Adam falsely believes that his desire doesn't guarantee that he kill Bennie by *t3*. As Pereboom points out, because Adam is ignorant of the fact that his desire *does* guarantee he kill Bennie by *t3*, my view does not entail that Adam has committed to killing Bennie yet. Nor is that a plausible conclusion. Adam is uncommitted if he doesn't believe that his current motivations guarantee him to kill Bennie by *t3*. As long as Adam remains unsure, he can remain uncommitted up until the *t3* deadline. Pereboom contends that, because of *R*, leeway incompatibilists will nevertheless intuit

---

[342] In correspondence.

that Adam is blameless for deciding at *t1* to kill Bennie even though Adam has a robust alternative to this decision at *t1*.

It's unlikely that intuitions will play in Pereboom's favor. After all, Adam knows that his desire doesn't guarantee him to decide at *t1* to kill Bennie. And Adam falsely believes that he can decide *not* to kill Bennie at any future time. So Adam could have decided at *t1* to *not* kill Bennie.[343] That alternative is robust. Adam makes his immoral decision anyway. So it's intuitive that he's blameworthy for deciding at *t1* to kill Bennie. But it's not intuitive that he's blameworthy for deciding by *t3* to kill Bennie, as he has no robust alternative to that decision.

In sum, Pereboom's response to Ginet's criticism of Fankfurt cases faces a dilemma. Either Adam believes he's guaranteed to kill Bennie, or he doesn't. If he does, then it's intuitive that he's blameless. But as I've shown, Adam's true belief, together with his other psychological states, entails that he has already committed at *t0* to killing Bennie. Since Adam has no alternative to that commitment, *PRA* predicts that he's blameless for forming it. When he decides to act on that commitment is irrelevant. So a 'does believe' variation of the *Bomb* case is not a counterexample to *PRA*. If Adam does *not* believe that he's guaranteed to kill Bennie, then Pereboom is correct that Adam has a robust alternative to his deciding at *t1* to kill Bennie. However, this 'doesn't believe' variation of the *Bomb* case no longer elicits the intuition that Adam is blameless for deciding when he does. Nor does *PRA* entail that conclusion. Either way, Ginet's timing criticism in defense of *PRA* is vindicated against Pereboom's response. An agent

---

[343] Though, as Pereboom points out, Adam has no alternative to eventually deciding to kill Bennie.

cannot, then, be blameworthy for any ϕ-ing for which s/he had no alternative. So premise (4) of Haji's argument is false. *BRW* stands.

Bibliography

Arpaly, N. (2006). *Merit, Meaning, and Human Bondage; an Essay on Free Will*, Princeton University Press.

Audi, R. (2004). *The Good in the Right: A Theory of Intuition and Intrinsic Value*, Princeton.

Audi, R. (2008). "Intuition, Inference, and Rational Disagreement in Ethics", *Ethical Theory and Moral Practice*, 11, pp. 475-492.

Baier, K. (1966). "Moral Obligation", *American Philosophical Quarterly*, Vol. 3, No. 3, pp. 210-226.

Bell, M. (2009). "Anger, Virtue, and Oppression", in L. Tessman (ed.), *Feminist Ethics and Social and Political Philosophy: Theorizing the Non-Ideal*, pp. 168-83

Bennett, C. (2002). "The Varieties of Retributive Experience," *The Philosophical Quarterly*, 52, pp.145–63

Bennett. C. (2008). *The Apology Ritual: A Philosophical Theory of Punishment*. New York: Cambridge University Press.

Bodenhausen, G. Sheppard, L. Kramer, G. (1994). "Negative Affect and Social Judgment", *European Journal of Social Psychology*, Vol 24, pp. 45-62.

Bok, H. (1998). *Freedom and Responsibility*. (Princeton: Princeton University Press).

Boonin, D. (2008). *The Problem of Punishment*, Cambridge University Press.

Boxer, K. (2013). *Rethinking Responsibility*. Oxford.

Burgh, R. (1982). "Do the Guilty Deserve Punishment?" *The Journal of Philosophy*, Vol. 79, No. 4, pp. 193-210.

Capes, J. (2012). "Blameworthiness Without Wrongdoing", *Pacific Philosophical Quarterly* 93, pp. 417–437.

Carlsmith, K. (2006). "The Roles of Retribution and Utility in Determining Punishment", *Journal of Experimental Social Psychology* 42, pp. 437–451.

Carlsmith, K and Darley, J. (2008). "Psychological Aspects of Retributive Justice", *Advances in Experimental Social Psychology*, Volume 40, Chapter 4. pp. 193-236.

Carlsmith, K., Darley, J. and Robinson, P. (2002). "Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment", *Journal of Personality and Social Psychology,* Vol. 83, No. 2, 284–299

Clarke, R. (2005). "On an argument for the impossibility of moral responsibility", *Midwest Studies in Philosophy* 29: 13 – 24.

Clarke, R. (2013). "Some Theses about Desert"*. Philosophical Explorations*, Vol.16, No. 2, pp. 153–164.

Cohen, G. (2006). "Casting the First Stone: Who Can, and Who Can't, Condemn the Terrorists?." *Royal Institute of Philosophy Supplement*, 58, pp 113-136

Copp, D. (1997). "Defending the Principle of Alternate Possibilities: Blameworthiness and Moral Responsibility", *Noûs*, Vol. 31, No. 4, pp. 441-456

Cooper, J. (2007). *Cognitive Dissonance: Fifty Years of a Classic Theory*, (Sage Publications).

Crisp, R. (2006). "Hedonism Reconsidered", *Philosophy and Phenomenological Research* Vol. LXXII1, No. 3, November, 619-645.

Cupit, G. (1996). *Justice As Fittingness*. NewYork: Oxford University Press.
Dagger, R. (1993). "Playing Fair with Punishment" *Ethics*, Vol. 103, No. 3, pp. 473-488.

Darley, J. and Pittman, T. (2003). "ThePsychology of Compensatory and Retributive Justice", *Personality and Social Psychology Review*, Vol. 7, No. 4, 324–336.

Darwall, S. (2006). *The Second-Person Standpoint*, (Harvard University Press).

Darwall, S. (2009). "Authority and Second-Personal Reasons for Acting", *Morality, Authority, and Law: Essays in Second-Personal Ethics I*, pp. 135-150.

Dolinko, D. (1991). "Some Thoughts About Retributivism", *Ethics*, Vol. 101, No. 3 (April), pp. 537-559.

Douglass, F. (2009). *Narrative of the Life of Frederick Douglass* (Belknap).

Driver, J. (1992). "The Suberogatory", *Australasian Journal of Philosophy* Vol. 70, No. 3, pp. 286-295.

Ellsworth, P and Gross, S. (1994). "Hardening of the Attitudes: Americans' Views on the Death Penalty.", Journal of Social Issues 50, no. 2, pp. 19-52.

Everett, D. (2008). *Don't Sleep, There Are Snakes: Life and Language in the Amazonian Jungle*, (Vintage: New York).

Fendrich, J. (1967). "A Study of the Association among Verbal Attitudes, Commitment and Overt Behavior in Different Experimental Situations", *Social Forces*, Vol. 45, No. 3, pp. 347-355.

Feinberg, J, (1965). "The Expressive Function of Punishment", *The Monist*, Vol. 49, No. 3, Philosophy of Law, pp. 397-423.

Feinberg, J, (1970). "Justice and Personal Desert", *Doing and Deserving: Essays in the Theory of Responsibility*, Princeton, pp. 55-94.

Feinberg, J. (1984). *The Moral Limits of the Criminal Law, Volume 1: Harm to Others*. New York: Oxford University Press.

Feldman, F. (1995). "Adjusting Utility for Justice," *Philosophy and Phenomenological Research* 55, pp.567–585.

Fischer, J. (1994). *The Metaphysics of Free Will*. Oxford, Blackwell Publishers.

Fischer, J. (2006). *My Way: Essays on Moral Responsibility.* Oxford.

Fischer, J, and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. (Cambridge: Cambridge University Press).

FitzPatrick, W. (2008). "Moral Responsibility and Normative Ignorance: Answering a New Skeptical Challenge", *Ethics*, Vol. 118, No. 4, pp. 589-613.

Frankfurt, H. (1969). "Alternate Possibilities and Moral Responsibility", *The Journal of Philosophy*, Vol. 66, No. 23, pp. 829-839.

Gert, H., Radzik, L. and Hand, M. (1994). "Hampton on the Expressive Power of Punishment", *Journal of Social Philosophy*, Vol. 35 No. 1, Spring, pp. 79-90.

Gibbard, A. (1990). *Wise Choices, Apt Feelings; A Theory of Normative Judgment*, (Oxford University Press).

Ginet, C. (1990). *On Action*, (Cambridge University Press).

Ginet, C. (1996). "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing", *Noûs*, Vol. 30, Supplement: Philosophical Perspectives, 10, Metaphysics, pp. 403-417

Ginet, C. (2002). "Living without Free Will by Derk Pereboom" Review, *The Journal of Ethics*, Vol. 6, No. 3, pp. 305-309

Goldberg, J., Lerner, J., Tetlock, P. (1999). "Rage and Reason: The Psychology of the Intuitive Prosecutor", *European Journal of Social Psychology*. 29, 781-795.

Goldman, A (2007). "Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status", *Grazer Philosophische Studien* 74, pp. 1–26.

Griffin, J. (1989). *Well-Being: It's Meaning, Measurement, and Moral Importance*, Clarendon Press.

Haidt, J. (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment", *Psychological Review*, Vol. 108. No. 4, pp. 814-834

Haji, I. (1998). *Moral Appraisability: Puzzles, Proposals, and Perplexities*, Oxford.

Haji, I (2002). *Deontic Morality and Control*, Cambridge University Press.

Hanna, N. (2013). "Two Claims about Desert", *Pacific Philosophical Quarterly* 94, pp. 41-56.

Hampton, J. (1991). "A New Theory of Retribution", in *Liability and Responsibility: Essays in Law and Morals*, ed. Frey, R.G. and Morris, C. (Cambridge: Cambridge University Press)

Hampton, J. (1992). "Correcting Harms versus Righting Wrongs: The Goal of Retribution," *UCLA Law Review* 39, pp. 1659-1702

Hart, H. L. A. (1968). *Punishment and Responsibility: Essays in the Philosophy of Law* (Oxford University Press).

Hauser, M., Young, L., and Cushman, F. (2008). "Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions", *Moral Psychology Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, ed. Sinnott-Armstrong, W. (MIT Press), pp. 107-144.

Honderich, T. (1988). *A Theory of Determinism: The Mind, Neuroscience and Life-Hopes* (Oxford University Press).

Hurka, T. (2001). "The Common Structure of Virtue and Desert," *Ethics* 112, pp. 6-31.

Isen, I. (2008). "Some Ways in Which Positive Affect Influences Decision Making and Problem Solving" in *Handbook of Emotions*, ed. Lewis, M., Haviland-Jones, J., and Feldman-Barrett, L. (Guilford Press). pp. 548-573.

Kagan, S. (1989). *The Limits of Morality*, (Clarendon: Oxford University Press).

Kagan, S. (2003). "Comparative Desert," in S. Olsaretti (ed.) *Desert and Justice*. Oxford, pp. 93-122.

Kagan, S. (2009). "Well-Being as Enjoying the Good", *Philosophical Perspectives*, 23, pp. 253-272.

Kane, R. (1996). *The Significance of Free Will,* (Oxford University Press).

Kant, I. (1998). *Critique of Pure Reason*. Ed. Guyer, P. and Wood, A. (Cambridge).

Keltner, D., Ellsworth, P. C., & Edwards, K. (1993). "Beyond simple pessimism: Effects of sadness and anger on social perception." *Journal of Personality and Social Psychology*, 64, pp. 740–752.

Kleinig, J. (1971). "The Concept of Desert", *American Philosophical Quarterly* Vol. 8, No. 1, pp. 71-78.

Kleinig, J. (1991). "Punishment and Moral Seriousness", *Israel Law Review* 25, pp. 401-421.

Knobe, J. and Leiter, B. (2007). "The Case for Nietzschean Moral Psychology" in *Nietzsche and Morality*, ed. Leiter, J. and Sinhababu, N. (Clarendon: Oxford University Press).

Kornblith, H. (2015). "Naturalistic Defenses of Intuition" in *Experimental Philosophy, Rationalism, and Naturalism Rethinking philosophical method*, ed. Fischer, E and Collins, J. (Routledge), pp. 151-168.

Lenman, J. (2006). "Compatibilism and Contractualism: The Possibility of Moral Responsibility" *Ethics,* Vol. 117, No. 1, pp. 7-31.

Levy, N. (2009). "Culpable Ignorance and Moral Responsibility: A Reply to FitzPatrick", *Ethics*, Vol. 119, No. 4, pp. 729-741.

Levy, N. (2009). "Culpable Ignorance and Moral Responsibility: A Reply to FitzPatrick", *Ethics*, Vol. 119, No. 4, pp. 729-741

MacNamara, C. (2010). "Taking Demands out of Blame" in *Oxford Handbook on Blame*, (Oxford University Press), pp. 141-161.

McGeer, V. (2015). "Building a better theory of responsibility", *Philosophical Studies* 172 (10) pp. 2635-2649.

McKenna, M. (2012). *Conversation and Responsibility*, Oxford University Press.

McLeod, P. (1999). "Desert and Institutions", in L. Pojman and O. McLeod (eds) *What Do We Deserve?* NewYork: Oxford University Press, pp. 186-195.

Mill, J.S. (1872/2009), *Utilitarianism*, Oxford.

Miller, (1999). *Principles of Social Justice*, Harvard University Press.

Moore, M. (1987). "The Moral Worth of Retribution" in *Character, Responsibility, and the Emotions*, ed F. Schoeman (Cambridge: Cambridge University Press), pp. 179-218.

Moore, M. (1993). "Justifying Retributivism," *Israel Law Review* 2, pp.15–49.

Moriarty, J. (2003). "Against the Asymmetry of Desert," *Noûs* 37, pp.518–36.

Morris, H. (1968). "Persons and Punishment", *The Monist*, Vol. 52, No. 4, Human Rights, pp. 475-501.

Morris, S. (2015). "Vargas-Style Revisionism and the Problem of Retributivism", *Acta Analytica*, 30 (3), pp. 305-316.

Mundle, C. W. K. (1954). "Punishment and Desert", *The Philosophical Quarterly*, Vol. 4, No. 16, pp. 216-228

Murphy, J. (1971). "Three Mistakes about Retributivism", *Analysis*, Vol. 31, No. 5, pp. 166-169

Murphy, J. (1973). "Marxism and Retribution", *Philosophy & Public Affairs*, Vol. 2, No. 3, pp. 217-243.

Murphy, J.G. (1988). "Hatred: a Qualified Defense", in Murphie, J. G. and Hampton, J. *Forgiveness and Mercy*, (Cambridge, Cambridge University Press), pp. 88-110.

Murphy, J. (2007). "Legal Moralism and Retribution Revisited", *Criminal Law and Philosophy*, 1, pp. 5-20.

Nelkin, D. (2011). *Making Sense of Freedom & Responsibility*, Oxford.

Nichols, S. (2007). "After Incompatibilism: A Naturalistic Defense of the Reactive Attitudes", *Philosophical Perspectives*, 21, Philosophy of Mind, pp. 405-428.

Nichols, S. and Knobe, J. (2007). "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." *Nous* 43, pp. 663–85.

Nichols, S and Prinz, J. (2010). "Moral Emotions" in *The Moral Psychology Handbook*, ed. John M. Doris and the Moral Psychology Research Group, (Oxford University Press), pp. 111-146.

Nietzsche, F. (1966). *Beyond Good and Evil*, trans. W. Kaufmann (New York: Vintage).

Nietzsche, F. (1998). *On the Genealogy of Morality*, trans. Maudemarie Clark and Alan Swensen (Indianapolis: Hackett).

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*, Oxford.

Parfit, D. (1984). *Reasons and Persons*, Clarendon Press, Oxford.

Parfit, D. (2011). *On What Matters: Volume I* (Oxford).

Pereboom, D. (2001). *Living Without Free Will*, Cambridge University Press.

Pereboom, D. (2007). "Hard Incompatibilism" and "Response to Kane, Fischer, Vargas" in *Four Views on Free Will*, Blackwell, pp. 85-125, 191-203.

Pereboom, D. (2009). “Free Will, Love, and Anger,” *Ideas y Valores: evista de Colombiana de Filosofía* 141, pp. 169-189.

Pereboom, D. (2012). "Frankfurt Examples, Derivative Responsibility, and the Timing Objection", *Philosophical Issues* 22, pp. 298-315.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*, Oxford.

Primoratz, I. (1989). "Punishment as Language", *Philosophy*, Vol. 64, No. 248 , pp. 187-205.

Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of the Emotions*, (Oxford).

Quigley, B. and Tedeschi, J. (1996). "Mediating Effects of Blame Attributions on Feelings of Anger", *Personality and Social Psychology Bulletin*, 22(12), pp. 1280-1288

Rawls, J. (1991). *A Theory of Justice Revised Edition* (Cambridge: Belknap press of Harvard University Press.

Rawls, J. (2003). *Justice as Fairness: A Restatement* (Cambridge: Belknap of Harvard University Press, 2003).

Rosen, G. (2004). "Skepticism about Moral Responsibility", *Philosophical Perspectives*, 18, pp. 295-313.

Rosen, G. (2015). "The Alethic Conception of Moral Responsibility" in *The Nature of Moral Responsibility: New Essays,* ed. Clarke, R., McKenna, M., Smith, A., Oxford, pp. 65-88.

Ross, W.D. (1930). *The Right and the Good*, Oxford.

Scanlon, T. T(1988). "The Significance of Choice", in S. McMurrin (ed.), *The Tanner Lectures on Human Values*, vol. viii (Salt Lake City: University of Utah Press), pp. 151-216.

Scanlon, T. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*, Belknap.

Scanlon, T. (2013). "Giving Desert its Due", *Philosophical Explorations,* Volume 16, Issue 2, pp. 101- 116.

Scheffler, S. (2003). "Distributive Justice and Economic Desert" in *Desert and Justice*, ed Serena Olsaretti, (Oxford: Oxford University Press), pp. 69-92.

Scher, S.J. and Cooper , J. (1989) "Motivational Basis of Dissonance: the Singular Role of Behavioral Consequences," *Journal of Personality and Social Psychology*, 56, pp. 899-906.

Schmidtz, D. (2002). "How to Deserve," *Political Theory* 30, pp.774–799.

Shabo, S. (2012a). "Where Love and Resentment Meet: Strawson's Intrapersonal Defense of Compatibilism", *Philosophical Review*, Volume 121, Number 1: 95-124.

Shabo, S. (2012b). "Compatibilism and Moral Claimancy: An Intermediate Path to Appropriate Blame", *Philosophy and Phenomenological Research*, pp. 158-186.

Scheffler, S. (1992). "Responsibility, Reactive Attitudes, and Liberalism in Philosophy and Politics", *Philosophy & Public Affairs*, Vol. 21, No. 4, pp. 299-323.

Scheffler, S. (2003). "Distributive Justice and Economic Desert" in *Desert and Justice*, ed Serena Olsaretti, (Oxford: Oxford University Press), pp. 69-92.

Shafer-Landau, R. (2000). "Retributivism and Desert", *Pacific Philosophical Quarterly* 81, 189–214.

Sher, G. (1987). *Desert*, Princeton, NJ: Princeton University Press.

Sher, G. (1997). *Approximate Justice: Studies in Non-Ideal Theory*, Rowman & Littlefield.

Sher, G. (2006). *In Praise of Blame*, Oxford.

Shoemaker, D. (2007). "Moral Address, Moral Responsibility, and the Boundaries of the Moral Community", *Ethics*, Vol. 118, No. 1, Symposium on Stephen Darwall's *The Second-Person Standpoint*, pp. 70-108.

Shoemaker, D. (2011). "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility", *Ethics* 121, pp. 602-632.
Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford.

Sinnott-Armstrong, W. (2006). "Moral Intuitionism Meets Empirical Psychology" in *Metaethics after Moore*, ed Terry Horgan and Mark Timmons, Oxford University Press, pp. 339-40.

Smilansky, S. (2002). "Free Will, Fundamental Dualism, and the Centrality of Illusion" in *The Oxford Handbook of Free Will*, ed Robert Kane, (Oxford: Oxford University Press).

Smith, A. (2007). "On Being Responsible and Holding Responsible", *The Journal of Ethics*, Vol. 11, No. 4, pp. 465-484.

Skorupski, J. (1999). *Ethical Explorations*. Oxford: Oxford University Press.

Smart, J.J.C. (1961). "Free-Will, Praise and Blame", *Mind*, New Series, Vol. 70, No. 279, pp. 291-306.

Steele, C.M. (1988) "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self," in L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, vol. 21. (San Diego, CA: Academic Press), pp. 261-302.

Steele, C.M. and Liu, T .J. (1983) "Dissonance Processes as Self-Affirmation," *Journal of Personality and Social Psychology* 45, pp. 5-19.

Stern, L. (1974). "Freedom, Blame, and Moral Community", *The Journal of Philosophy*, Vol. 71, No. 3, pp. 72-84

Strawson, G. (1986). *Freedom and Belief*, Oxford University Press.

Strawson, G. (1994). "The Impossibility of Moral Responsibility", *Philosophical Studies*, Vol. 75, No. 1/2, pp. 5-24.

Strawson, G. (2002). "The bounds of freedom" in *The Oxford handbook of free will*, ed. Robert Kane (Oxford: Oxford University Press), pp. 441-460.

Strawson, G. (2008). "On Freedom and Resentment" in *Free Will and Reactive Attitudes: Perspectives on Peter Strawson's 'Freedom and Resentment'*, ed. Michael McKenna and Paul Russell (Ashgate), pp. 85-114

Strawson, P. (1982). "Freedom and Resentment", *Free Will* (1st edn) (Oxford University Press), pp. 59–80. Reprinted in *Free Will and Reactive Attitudes: Perspectives on Peter Strawson's 'Freedom and Resentment'*, ed. Michael McKenna and Paul Russell, (Ashgate 2008), pp. 19-36.

Vargas, M. (2013). *Building Better Beings*. (Oxford University Press).

Vargas, M. (2015). "Desert, Responsibility, and Justification: a Reply to Doris, McGeer, and Robinson", *Philosophical Studies* 172 (10), pp. 2659-2678.

Vidmar, N. (2001). "Retribution and revenge", In V. L. Hamilton (Ed.), *Handbook of Justice Research in Law* (pp. 31–63). New Haven, CT: Yale University Press.

Villhauer, B. (2009). "Free Will Skepticism and Personhood as a Desert Base", *Canadian Journal of Philosophy* 39(3), pp. 489-511.

Vilhauer, B. (2013). "Persons, Punishment, and Free Will Skepticism", *Philosophical Studies* 162, pp. 143–163.

Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*, Harvard.

Wallace, R.J. (2010). "Hypocrisy, Moral Address, and the Equal Standing of Persons", *Philosophy & Public Affairs* 38, no. 4, pp.307-341.

Waller, B. (2011). *Against Moral Responsibility*, (MIT Press).

Waller, B. (2015). *The Stubborn System of Moral Responsibility*, (MIT Press).

Walton, K. (1978). "Fearing Fictions," *The Journal of Philosophy* 75, pp. 5-27.

Watson, G. (1987). "Responsibility and the Limits of Evil" in *Responsibility, Character, and the Emotions* ed. Schoeman, Ferdinand (Cambridge), pp. 256-286.

Watson, G. (1996). "Two Faces of Responsibility", *Philosophical Topics* Vol. 24 No.2, pp. 227-248.

Widerker, D. (1991). "Frankfurt on 'Ought Implies Can' and Alternative Possibilities", *Analysis*, Vol. 51, No. 4, pp. 222-224

Zimmerman, M. (1988). *An Essay on Moral Responsibility*, Rowman and Littlefield.

Zimmerman, M. (1997). "A Plea for Excuses", *American Philosophical Quarterly*, Vol. 34, No. 2, pp. 229-243