UNIVERSITY OF CALIFORNIA

Los Angeles

On the Possibility and Permissibility of Interpersonal Punishment

A dissertation submitted in partial satisfaction of the

Requirements for the degree of Doctor of Philosophy

In Philosophy

by

Laura Gillespie

2017

ABSTRACT OF DISSERTATION

On the Possibility and Permissibility

of Interpersonal Punishment

by

Laura Gillespie

Doctor of Philosophy in Philosophy

University of California, Los Angeles, 2017

Professor Seana Shiffrin, Chair

In the dissertation, I consider the permissibility of a familiar set of responses to wrongdoing in our interpersonal relationships—those responses that constitute the imposition of some cost upon the wrongdoer. Some of these responses are, I argue, properly considered punishing, and some of these instances of punishing are in turn permissible. Punishment as I understand it is a broad phenomenon, common in and to all human relationships, and not exclusively or even primarily the domain of the state. Personal interactions expressive of wrong-reactive attitudes like disappointment, anger, and guilt will sometimes constitute punishment so understood. I consider childhood punishment, self-punishment, and punishment between friends, concluding that punishment in the context of our personal relationships may sometimes be appropriate where undertaken not for the sake of deterrence nor of retributive justice, but for the sake of the aims constitutive of the relationship in which it occurs.

The dissertation of Laura Gillespie is approved.

Sharon Dolovich

Barbara Herman

Alexander Jacob Julius

Seana Shiffrin, Committee Chair

University of California, Los Angeles

2017

For Gram.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

For the past several years I have had the honor of being allowed to walk casually into the offices of a roster of legends, to make a little small talk, and then, incredibly, to discuss my own work with them—to hear what they have to say about it. It's astonishing, really. My thanks are due, first of all, to the legends in question: my committee, and especially its chair, Seana Shiffrin. Seana has given more to me and to this project than I could ever repay. She is a lion, and I sincerely thank her for the honor of being her student. There isn't a single member of my committee, finally, whose attention to my work hasn't had an enormous impact on its final shape, or on my sense of who I'd like to be as a philosopher. The impossible sensitivity, insight, care, and good humor they unfailingly bring to their own work, they've also brought to mine. If I live to be a productive philosopher of 100, I'll still be coming back to these conversations and learning something new from them. Thank you Barbara Herman, Pamela Hieronymi, AJ Julius, Sharon Dolovich, and Erin Kelly. My cup runneth over.

Thanks also to the wide variety of friends and colleagues in the profession, and especially in the UCLA philosophy department, who have read and commented on drafts over the years, and generally made the academy into a space that feels sane and habitable for me. Particular thanks to the participants in UCLA's Ethics Workshop (with special gratitude to Greg Antill and Adam Masters), and to Pamela Hieronymi for running it; to my comrades from day one, Jonathan Gingerich and Lauren Schaeffer; to Amelia Acker, of whom I am permanently in awe; and—finally, forever—to the thread.

Thanks to my friends and family, who tolerated several years of relative neglect and showed up anyway, over and over again, to absorb some of the anxious vibrations and remind me of who I am. Particular thanks to Leah Gillespie, my sister and my person forever; to Barbara

and T.C. Gillespie—you're weird parents, but I love you; to Rosie Wagner (there aren't words, honestly); and to my guy, Alex Delyle, who kept it all going.

This dissertation was supported by a Dissertation Year Fellowship from the Graduate Division of the University of California, Los Angeles.

CURRICULUM VITAE

EDUCATION

M.A. in philosophy, Tufts University, Medford, Massachusetts 2006-2010

B.A. in philosophy, Simmons College, Boston, Massachusetts 2000-2004

GRADUATE AWARDS & HONORS

2016–2017	UCLA Dissertation Year Fellowship
2015–2016	Departmental Distinguished Teaching Award
2014–2015	Dean's Humanities Fellowship
Summer 2014	Mellon Fellowship
Summer 2012	Graduate Student Research Mentorship with A.J. Julius
Summer 2011	Mellon Fellowship
2010–2011	Dean's Humanities Fellowship

PRESENTATIONS & CONFERENCE PARTICIPATION

Summer 2016	"Between Friends", Gothenburg Responsibility Project Workshop, University of Gothenburg
Summer 2016	*Session Chair*, Athena in Action, Princeton University.
Spring 2016	*Session Chair*, Pacific APA, San Francisco.
Spring 2016	"Love as Reverence and Revelation: Mad Love in Plato's *Phaedrus*", Comparative Literature Graduate Student Conference, UCLA.
Fall 2015	"Between Friends: Interpersonal Punishment in Relationships of Equality", Seminar: Mutual Recognition with A.J. Julius, UCLA.
Fall 2015	"Between Friends", Ethics Workshop, UCLA.
Summer 2015	*Session Chair*, Bellingham Summer Philosophy Conference, University of Washington, Bellingham.
Summer 2015	"The Participant View: A Theory of Childhood Punishment", Rocky Mountain Ethics Conference, University of Colorado Boulder. *(Selections published in invited post to* What's Wrong?*, the blog of UC Boulder's Center for Values and Social Policy.)*
Spring 2015	"Punishment Among Equals", Albritton Talk, UCLA.
Winter 2015	"Three Conditions on Punishment", Ethics Workshop, UCLA.
Fall 2014	"The Participant View: A Theory of Childhood Punishment", Ethics Writing Group, USC.
Winter 2014	"Against Shaming Punishments", Ethics Workshop, UCL

Chapter 1. The Possibility of Interpersonal Punishment

The imposition of hard treatment in response to wrongdoing is a familiar feature of human life. A significant portion of us will at one time or another suffer formal, criminal punishment at the hands of the state.[1] [2] Virtually all of us will experience punishment in informal, personal contexts. As children we suffer the loss of certain pleasures and privileges at the hands of caregivers when we misbehave. As adults we stand as objects of our own recriminations and deny ourselves pleasures and rewards for our own perceived failures, and we condemn and withhold from one another as we ourselves feel wronged. We may also impose or experience forms of punishment that are public in nature, but distinct from that which the state may impose—setting out, e.g., either individually or collectively to shame or dislodge from her position a public figure whose behavior we determine to be hypocritical, cruel, or unjust. In whatever domain of human life one finds oneself there will be moral failure, and hard-treatment in some set of its myriad forms will always be among the familiar responses to it.

While there is ostensibly a vast literature on the philosophy of punishment, it is better characterized as a literature on the philosophy of state criminal punishment. The central concern

---

[1] According to the US Bureau of Justice Statistics roughly 1 in 110 Americans at any given time are incarcerated in our state prisons, federal prisons, or county jails, and almost 1 in 37 is under some form of correctional supervision (Kaeble 2015).

[2] Throughout the course of this chapter and the dissertation as a whole, I will draw out a contrast between punishment as meted out by the state and punishment (according to my understanding) as meted out in the context of our personal relationship. When I speak of state punishment in this context, I mean, in particular, criminal punishment. It issues not only criminal punishments, but levies civil penalties and awards punitive damages. Employees of our state-run schools sometimes punish students. I do not mean to deny that these are, in fact, instances of state punishment, or to imply that the theory of punishment I will offer has nothing to say about them. For current purposes, though, I am putting these cases to the side, and when I speak of state punishment, I mean to speak, in particular, of hard-treatment meted out for violations of the criminal law.

of philosophers writing about punishment has largely been to capture and provide adequate justification for the state practice of punishing law-breakers, particularly where punishment takes the form of incarceration or killing. Some have gone so far as to exclude hard-treatment imposed by an agent other than the state from the very definition of punishment. On this understanding, anything with punishment's general character occurring before the advent or outside the purview of the modern state is only punishment in a primitive or partial sense. While not all philosophers of punishment will be inclined to exclude personal (as opposed to institutional) responses to wrongdoing from the strict definition of punishment, few have treated interpersonal punishment as a subject worthy of serious interest.

The near exclusive focus on deprivations and burdens imposed by the state for the violation of criminal statutes would seem to reflect a general sense that other forms of punishment (if indeed there are any) fail to raise any interesting philosophical questions—interesting either in general or, at least, to those who seek to understand the practice of state punishment in particular. There are at least three sorts of reasons why this might be. Perhaps (1) all or most other kinds of punishment are straightforwardly impermissible in the context of the modern state and its "monopoly on violence".[3] Perhaps (2) other kinds of punishment can be explained and justified by some extension of existing theories of state punishment; or perhaps (3) other, more personal forms of punishment are so different from state punishment in character or import that they need not and perhaps ought not inhabit the same literature.

[3] The notion of the state as an institution that holds the exclusive right to use or threaten physical force against residents of its territory goes back at least to Hobbes, but was first characterized as a defining feature of the state by sociologist Max Weber. Weber argued that part of what it is to be a state is to successfully hold such a "monopoly on the legitimate use of physical force," primarily by means of a police and military force (Weber 1922, p23).

We should dispute all three of these claims. Consider the first: that punishment cannot occur permissibly outside of legal and institutional contexts. The notion that interpersonal punishment would violate the state's exclusive authority to use, threaten, or authorize violence assumes that all punishment is violent, or, at least, that it involves some threat of force. Our common experience tells against this claim. For one thing, it is not true that the state alone is entitled to use or threaten physical force, even in punishing. This certainly looks to be true in the case of childhood punishment, where even non-corporal forms of punishment like time-outs and losses of privileges often involve physically moving, or giving up some object, where this is compulsory. While not uncontroversial in some corners of the world, the practice of punishing children is nonetheless still widely held to be permissible or even morally required of those responsible for a child's general welfare.

What's more, though, not all punishment involves physical force, or even the distant threat of it. In the course of adult life we will be subjected to a range of interpersonal treatment that we might plausibly construed as punishing—treatment may be painful or otherwise unwanted, but neither coercive nor violent. In the normal course of a human life we will all at times impose more or less subtle forms of sanction upon ourselves and others in response to alleged or perceived wrongs—wrongs that may but generally will not rise to the level of law breaking, but violate, rather, the rules of friendship, marriage, or neighborliness. Granted, many such impositions may be misguided or otherwise objectionable. The withholding of normal affection or attention, for instance, as a means of doing harm to those we feel have betrayed us or let us down may often turn out to be petty, vindictive, patronizing, manipulative, or passive-aggressive. But many real-world instances of criminal punishment turn out likewise to be objectionably vindictive, patronizing, manipulative, and so on, and we do not (or at least ought

not), in this cases, infer from the fact that punishment has very often gone badly wrong in practice to the conclusion it can never go right. Nor can we take this to settle the question of what possible place or value the state practice or punishment criminals, properly done, might have. Establishing whether or not anything we want to call punishment actually turns out to serve some valuable end, and whether such treatment turns out to be a permissible means of serving it, requires further investigation. This is as true of interpersonal punishment as it is of criminal punishment by the state.

If there are any such cases of permissible interpersonal punishment, we should also doubt that (2) our standard theories of legal punishment can be easily extended to characterize or provide us with proper conditions on their permissibility. We should doubt, for example, that there is any straightforward way of extending a retributive theory of punishment to justify the punishment of children by a caregiver, especially where it occurs in very early childhood.[4] Likewise, we would be hard pressed to explain any permission we may have to shame a hypocritical televangelist in terms of an individual or collective right to security or self-

[4] I understand retributivism to be the view that wrongdoers deserve to suffer some punishment proportionate to the wrong they have done, and that where there is some legitimate agent of punishment, it is a good in itself that this agent so punishes the wrongdoer. On this view, "punishment is justified by the desert of the offender" (Murphy 2007). One might be unwilling to think that childhood punishment could be justified in this way, either because the wrongs in question are generally not serious enough to warrant punishment, or because children are not yet full-fledged agents of the sort that can deserve something in this way. There are many possible standards for who counts as an agent and proper object of punishment, but young children and the seriously mentally ill, at least, are generally thought to be on the list of persons who cannot be said to meet these standards. One might also worry that there is something about the nature of the parent-child relationship itself that makes childhood punishment a bad candidate for a retributive analysis—that "it would be perverse," given the caring and tutelary nature of the relationship, "if the parents were generally to punish primarily from motives of retributive justice" (Morris 1981, p267).

protection.[5] Deterrence and moral education accounts, too, seem particularly misplaced in the context of our personal relationships, where paternalism and bare attempts at manipulating behavior seem particularly problematic. The notion that one might punish a friend as a means of manipulating her behavior or "teaching her a lesson" is a deeply distasteful one. If there are indeed instances of permissible interpersonal punishment common to human life, they resist easy analysis in terms of any one of our existing theories of state punishment.

That interpersonal punishment may not be fully defensible or even interpretable as such by appeal to our standard theories of state punishment should come as no surprise. The relationship between a citizen and the state is very different in character from the relationship each of us has to ourselves, to our friends, children, spouses, and even to those persons who act as our representatives in government, or as authority figures in our chosen moral communities. The state's relationship to a citizen is impartial, and state punishment must be administered by an impartial judiciary. But interpersonal punishment, if there is indeed such a thing, will be partial. Though not untethered from concerns about fairness, justice, and commensurability, it occurs in the context of our personal relationships, and may be expressive of a range of emotions and attitudes beyond those appropriately or even conceivably expressed by the state, as impartial representative, to a law-breaker. The state, by its nature, cannot withhold affection or approval in

[5] Alternatives to strictly retributivist accounts of punishment have focused especially on the rights of people against certain kinds of harms, and the corresponding responsibility of a just state to provide protection against those harms, punishing those who intentionally cause them insofar as a policy of punishing such wrongdoers has a deterrent effect (see Hart 1968, Rawls 1955, Kelly 2009). In the case of non-legal punishment we are dealing with wrongs that do not violate the rights of citizens as such, insofar as there will or ought to be a law against such an act. But even in the context of relationships other than that of a state to its citizens, the general or special deterrent effect of punishment might constitute a reason to punish. It seems clear, though, that when we shame someone it is at least in some measure because we feel they ought to be ashamed of themselves.

the way a friend or parent can, as the friend, by her nature, cannot invalidate one's legal right to vote.

The final claim we should deny is that (3) punishment in interpersonal contexts has nothing of central import to teach us about the nature and possible legitimacy of state punishment. Attending to punishment as it occurs across the full range of human experience has the potential to enrich the pool of resources from which we as philosophers draw in theorizing state punishment. First, it allows us to take a conceptual step back, generalizing from punishment as it occurs in a range of contexts to establish what we can about the deep architecture of this particular form of redress. When we can say with greater clarity what punishment is (and I think we can), we provide grounds for a clearer discussion of what punishment *does* in any given context—state or otherwise. Second, when we attend to punishment as it occurs across a range of contexts we gain a wider sense of the possible forms punishment can take and the different kinds of value it might serve. Not all of what we learn in one context will be easily importable into another. Punishment for the sake of moral education may be suitably imposed by the parent upon the child, but may strike us as more problematic when so imposed by the state upon its citizens. Punishment for the sake of deterring bad behavior may be suitably imposed upon the law-breaker by the state, while being wholly inappropriate in the context of a friendship. Still, when we turn our attention to punishment in these under-theorized interpersonal contexts, we may discover that punishment can take forms and serve ends other than those we have traditionally practiced or considered, but which may be worthy of consideration or practice.

One of the central features of punishment that we may miss when we adopt a myopic focus on state punishment is the centrality of relationships, both in terms of punishment's form and, finally, its point. Punishment is a particular form of treatment (*hard* treatment), imposed by

one agent upon another in response to wrongdoing. These agents— punisher and punished— are always in some form of relationship. This relationship may be formal or informal; it may be more or less intimate; it may be very particular and idiosyncratic, or quite general and distant. Even to be simply fellow human beings is to stand in a kind of norm-governed relationship to one another.[6] What constitutes a wrong in any particular case itself depends on the terms of the relationship. When I wrong you, I wrong you in some capacity—*as a person*, or *as a colleague*, or *as a friend*. Punishment is a response to wrong, imposed by one member of this relationship (the alleged norm violator) against another. So though the word "relationship" may not appear explicitly in any of the going definitions of punishment and is rarely used in our discussions on the matter, the very form of the act assumes relationships as a background condition. When we are talking about punishment we rely implicitly on the notion of a background relationship.

Whether or not any particular instance or variety of punishing turns out to be permissible is something else we cannot settle without thinking about relationships, and here we must look to particulars. The question of what constitutes a proper response to wrongdoing, like the question of what constitutes wrongdoing in the first place, is something we can only judge against the terms of the particular relationship in which it takes place. There is a version of this point that has been made or at least suggested by several prominent philosophers in their work on state punishment. These philosophers offer arguments for punishment that draw on a more general account of the state's authority, and of the liberal democratic values that they take to underwrite

[6] I have in mind here something like what T.M. Scanlon has called "the default moral relationship", which consists in "the kind of mutual concern," or "mutual regard and forbearance…that, ideally we all have toward other rational beings" (Scanlon 2008, p140-141). To stand in this sort of relationship to another person is just to owe them the basic forms of consideration that we owe to all other persons as such.

that authority.[7] If we want to know whether or not punishment of some particular form or other is an appropriate state response to law-breaking, we have to know something about the nature and limits of state authority as a general matter. If we want to establish that punishment of some particular form falls short of an ideal of state action with respect to citizens, we had better have some account of what the relationship between citizen and state ought to be. Absent such an account, we have no way of establishing the state's authority to pursue whatever valuable end punishment serves, whether as a general matter or by some particular means.

There is a more general point to be made here, which holds for punishment across all domains: For any two parties *A* and *B*, whether or not *A*'s punishing response to *B*'s wrongdoing is appropriate will depend not only upon the value that the imposition aims to serve, but also on the terms of the relationship the two parties stand in to one another. It is *as* a member or *citizen* of that relationship that *A* punishes. The parent punishes *qua* parent, and whatever right and responsibility she has to do so will be grounded in the more general set of rights and responsibilities she has in that role. So too in the case of colleagues, neighbors, lovers, teammates, and friends. Each of these is a particular kind of norm-governed relationship, and to be a member of such a relationship is to have a particular set of rights and responsibilities as such. To be in a relationship is, in other words, to have a particular form of authority. This is true even in relationships of equality, where the authority is non-hierarchical. You have, as my friend, the right to ask certain things of or about me that a stranger would not have. You are, then, as a friend who knows and cares about me, a kind of authority *on* me, and have some authority *over* me, too (though not the kind that could entitle you to force my compliance, should I choose to

[7] See Dolovich 2004, Duff 1986, Feinberg 1984-1988, Hampton 1994, Morris 1981, and Murphy 1973.

flout your friendly authority in any particular instance). Whether that authority might ever include the right to punish is something we can only settle by appeal to some account of the norms governing that particular kind of relationship.

One of the primary advantages of looking at punishment as it occurs in interpersonal contexts is that it foregrounds relationships in our conception of punishment and of punishment's possible value—a move that can help us to better understand the (limited) place and value of punishment in human life more generally. Over the course of this dissertation, I consider a range of philosophical questions about interpersonal punishment, including both the question of (first) whether there is any such thing, and (finally) what relationship interpersonal punishment (if and as it goes on across a range of interpersonal contexts) has to institutional punishment—in particular, the punishment of citizens by the state. I will advance three central claims. The first is that *punishment is a broad phenomenon, common to all agential relationships and not the exclusive nor even primarily the domain of the state.* The form of treatment of which state punishment is an instance includes a wider swath of our ordinary ways of treating one another than one might have thought. In the final part of this chapter I will establish that there are three (as opposed to the traditional five) necessary and sufficient conditions on punishment, and that some of the ways that we treat our children, our friends, and ourselves in response to wrongdoing meet this conditions.

The second central claim I will advance is that *punishment as it occurs in the domain of the personal raises special problems not easily addressed or even fully formulable in terms of any of the standard philosophical theories of punishment* (i.e. retributive, deterrence, or moral education theories). I have already touched on the uneasy fit between the picture of the aims and motives proper, on these theories, to punishment, and the motives and aims proper to parenthood,

friendship or a healthy self-regard. In chapters 2, 3, and 5, I will discuss this mismatch in the context of each of these three roles, respectively. More generally, though, we can see that these views of punishment, tailored to the theorizing of state action, condition permissibility on the punishing agent constituting the sort of impartial authority that the person operating in her capacity as friend, neighbor, or parent is certainly not.

The third central claim I will advance is that *a limited defense of punishment in a range of personal relationships—including relationships of equality, like friendship— is possible*. I have already suggested that a successful defense depends on first expanding in a principled way our understanding of what counts as punishment, then establishing the constitutive aim of the relevant relationship and the more particular obligations it generates. It will depend, too, I will argue, on locating the motive for punishment in a commitment to the relationship itself. Following this formula, I offer two relationship-centered accounts of interpersonal punishment: a theory of childhood punishment, and a theory of punishment in friendship, understood broadly to encompass most of our personal relationships with other morally mature agents. When we punish one another permissibly, if we ever do, it will be for the sake of neither deterrence nor vengeance, self-interest nor the demands of impartial morality, but for the sake of the particular relationship in which it takes place. In chapter 5 I consider the question of how we might apply this formula for theorizing interpersonal punishment in the intrapersonal case.

What will emerge over the course of theorizing punishment across this range of particular interpersonal contexts is a picture of punishment's deep architecture, and a justificatory strategy I have called "relationship-centered". Having developed this approach over the course of several chapters I return, bringing it to bear, finally, on the question of state punishment. State punishment, too, I argue, should be understood as taking place in the context of a particular kind

of relationship. The practice of state punishment is distinct in some ways but nonetheless continuous with the whole range of punishment practices common in human life, and, like these other practices, mustering the theoretical resources necessary for justifying state punishment will require us to attend to the aims and contours constitutive of the kind of relationship holding between punisher and punished. Some of us will not be in the habit of thinking about ourselves as having a "relationship" to the state, as we are accustomed to doing when thinking or speaking of the interactions that take place in our personal lives. It is, though, a kind of relationship, and thinking of it in these terms can, I will argue, help to generate further resources for a fuller characterization and more satisfying justification of state punishment. In theorizing interpersonal punishment, thus drawing out the relational dimension of the practice, we emerge with a theory that provides some novel and perhaps more satisfying answers to questions about state punishment.

In the rest of this chapter I will lay the groundwork for this theory of punishment by setting out first to establish a view about what punishment, as a general category, is, and why we should think that punishment is a relatively common form of treatment. The foundational idea of this dissertation is that punishment goes on regularly in various interpersonal contexts, and that we can learn something important about the concept of punishment and punishment's possible justification by considering these forms of interpersonal punishment. To entertain this idea, one must be willing to adopt the broad sort of understanding of punishment that I am about to set out. This understanding will constitute a sharp departure from the views standardly represented in the literature. I will devote a fair amount of time here to laying out the standard view, which I reject, then laying out an alternate view and making a case for it.

My reasons for rejecting the more restricted definition, though, are not simply or even primarily that it would exclude the class of cases I mean to consider. The more restricted definition, I will argue, has at least two important short-comings, which the proposed alternative avoids: First, it is both under- and over-inclusive in how it picks out the relevant cases, even if we only care about punishment by the state. Second, it builds permissibility conditions into the definition itself, rendering *punishment* and *permissible punishment* indistinguishable, and *impermissible* punishment no punishment at all. The standard, narrow definition begins from a concern to analyze what a particular, historically contingent set of language users mean by the word "punishment", rather than settling anything about what punishment fundamentally is. In the process, I argue, they fail on both counts, offering a definition of punishment that commits us to thinking that certain familiar uses of the term "punishment" are not really punishment at all, while simultaneously failing to pick out all and only those features distinctive of punishment as a class of action. Punishment, I will argue, is *the intentional imposition of hard-treatment in response to wrongdoing*. It is this form of treatment as such that raises the distinctive kind of justificatory problem with which philosophers of punishment have been concerned: *What is it about wrongdoing that licenses the intentional imposition of hard-treatment*? Where a form of treatment raises this distinctive justificatory problem, we gain nothing by withholding the label "punishment".

I then turn in the remainder of the dissertation to demonstrating what we have to lose by withholding it—namely, a useful and unified treatment of a class of human action with a common structure, which will include a general strategy of justification with application across the whole range of cases—including those of state punishment.

Two Definitions of Punishment

Though interest in the question of what might constitute a set of necessary and sufficient conditions on punishment has waned, there was a period of several years from the mid-1950's through the early-1970's when philosophers were quite concerned with it.[8] No entirely satisfying consensus ever emerged, and any serious attempt to establish such conditions has been largely abandoned. There was, though, finally, something of an uneasy agreement amongst at least one prominent set of voices, who shared a sense that the central or paradigm cases of punishment (perhaps the only cases of punishment, strictly speaking) in our modern context are those in which the punisher sits in a position of institutional authority.

In 1954 Anthony Flew proposed five conditions on punishment, subsequently adopted (with slight adjustments) and popularized by H.L.A. Hart and S.I. Benn.[9] These criteria are That it be "an evil, an unpleasantness, to the victim";[10]

(i) That it be for an offense;
(ii) That it be of the offender;
(iii) That it be "the work of personal agencies";[11]
(iv) That it be imposed in virtue of some authority, conferred through or by the institutions against

What Flew and company mean to capture by these conditions is what we typically mean when we say "punishment". Each condition allegedly captures something about the "genuine…meaning of 'punishment' in its standard sense."[12] Condition (i) tells us that punishment is a kind of harm done to the person punished. Flew is careful to note that on our

[8] See especially Flew 1954, Mabbott 1955, Benn 1958, Quinton 1959, Hart 1960, Armstrong 1961, McCloskey 1962, McPherson 1967, and Marshall 1972.
[9] Flew 1954, Hart 1960, Benn 1958.

[10] Flew 1954, p293.

[11] *Ibid*. p294.

[12] Scheid 1980, p485.

current understanding, this harm needn't involve the imposition of the physical pain of, e.g., flogging. The point here is one about the evolution of the word: While some idea of physical suffering may once have been "an essential part of the meaning of the word," it does not seem to be now.[13] Conditions (ii) and (iii) tell us that whatever the harm suffered might be, it must be a response to wrongdoing, imposed upon the wrongdoer. Condition (iv) says that "punishment must be the work of personal agencies," meaning that it cannot be a mere act of nature, but must be something intentionally imposed—an *act* issuing from some *actor*. While we may call the wind punishing, or conceive of some accident befalling us as a punishment for past wrongs, these uses must be figurative.[14] Finally, condition (iv) specifies that the agent in question must be acting on some form of institutional authority, where the offense to which her imposition responds constitutes some violation of the rules laid down by the authority-granting institution. While Flew himself did not specify which of these conditions he took to be either necessary or jointly sufficient, Benn and Hart, along with most others, seem to have taken them to be each necessary and jointly sufficient for punishment. Where all are met, they claim, we have punishment in its fullest sense.

Where controversy arose, it typically centered on conditions (iii) and (v). These conditions seem to cause problems if construed as strictly necessary. If it were a necessary condition on punishment that it be imposed on an actual offender, then propositions like "an innocent man has been punished" would turn out to be self-contradictory or even nonsensical. If it were a necessary condition on punishment that it be imposed on proper authority, then

[13] Flew 1954, p293.

[14] An exception that Flew notes are those cases in which we are claiming to have been punished by an interventionist God. Where such claims seem to be claims about literal as opposed to mere figurative punishment, it will be because the conception of God is of God as an agent—an actor, acting on the basis of reasons, and thereby capable of the purposive imposition of harm.

propositions like "you had no right to punish her" might run into similar problems. These propositions, however, are neither nonsensical nor contradictory. We know all too well what it is for an innocent man to be punished, or for the unauthorized mob to take justice into its own hands. It is Flew and company who are concerned to ground our understanding of punishment in the analysis of a set of linguistic conventions, and yet their definition rules out as punishment a class of treatment that those same conventions treat as, in fact, instances of punishing. And there is, moreover, significant cost to grounding our understanding in this way. We lose the idea of punishing the innocent as a distinct form of wrong, and, more generally, lose the traction we would need in order to properly distinguish between punishment and *permissible* punishment.

These counterexamples suggest an obvious problem with these conditions, which is that they seem to be conditions not on what constitutes punishment, but on what constitutes permissible punishment. There is nothing ill-formed about the sentence "an innocent man has been punished." The problem is in the activity itself. Punishing the innocent is something we typically judge to be neither wise nor just. That one is not permitted to punish without the proper authority to punish is even more obvious. It borders on the tautological. This, too, though, is a claim about the conditions on punishment's permissibility, not on the proper application of the concept.

Still, Flew and company were strongly inclined to hold on to both conditions. From Scheid (1980):

> "…[Conditions (iii) and (v) seem to be] a genuine part of the meaning of 'punishment' in its standard sense…[they] do not seem to represent contingent characteristics ("accidents") which merely happen to attend a high percentage of cases of punishment…If we consult our linguistic intuitions, it seems that criteria (iii) and (v) are somehow genuine parts of the meaning of 'punishment' in its standard sense, and they

have certainly much more than simply weak connotations which have become attached to the word." [15]

This somewhat vague but powerful sense that conditions (iii) and (v) are part of the very meaning of punishment is spelled out in a couple of ways. The first, bearing particularly on the status of condition (iii), springs from a sense that to drop the condition would constitutes a failure to respect the necessary conceptual tie between punishment and wrongdoing. Even in the case where we punish an innocent man, we must, after all, be punishing him for some crime that he is *alleged* or *perceived* to have committed. Without at least some tenuous connection between the imposition of the harm and some wrongdoing to which that imposition constitutes a response, it really would be strange to call the imposition a punishment.

To resolve this dilemma, philosophers tended to speak of punishment in a "strong"[16], "primary"[17], or "standard"[18] sense. For punishment in this sense, all five criteria must be met. There will then be weak, secondary, or non-standard cases that do not meet all five criteria, though these will have to be comparatively few.

---

[15] Scheid 1980, p485

[16] Armstrong's preferred distinction was between punishment in the "strong" sense, which met all five criteria, or a "weaker" sense. Punishment in the "weak" sense, though, does not turn out to be punishment at all, strictly speaking. The idea, rather, is that the word "punishment" can make sense in a weak sort of way when applied in cases that meet all five criteria. Armstrong's analogous example is the use of the term "kill" in the proposition "they half killed him" (Armstrong 1961, p479).

[17] This was the term that Flew himself preferred, conceding that the word "punishment" is often employed to describe cases that do no meet all five criteria. He considered these cases, though, to be relatively rare and therefore "non-standard" (Flew 1954, p292).

[18] This term is Scheid's. "As a solution to the problem," he says, "I introduce the notion of what I shall call a *reducible concept* (a nice awkward term); and my claim is that 'punishment' in its standard sense is a reducible concept. In any context in which 'punishment' would be taken in its standard sense, listeners will assume that all five criteria are met; but the speaker can reduce the sense of the word by denying or subtracting one or more of the criteria" (Scheid 1980, p461).

Why not, though, just weaken the condition, so that the object of punishment must be the *alleged* or *perceived* offender? Defenders of the narrow view take this to be inconsistent with their sense that the condition is "somehow [a] genuine part of the meaning of punishment." One reason they offer for thinking so is that the object or aim of punishment is *not* to sanction perceived or alleged wrongdoing. What punishment purports to do is sanction wrongdoing itself. If, then, there is no wrongdoing after all, it may seem that the punisher has not merely failed to punish in a way that achieves her objectives in punishing, but that she has failed to do the very thing she was trying to do. So perhaps her failure was not merely the failure to punish *permissibly*, but the failure to punish at all. So, too, with condition (v): What punishment purports to do is impose an *authorized* sanction—one the punisher is entitled to impose. If, then, one lacks the relevant form of authority, one has failed to do the thing at all. The party who has acted in a way that meets only conditions (i), (ii), and (iv) may not, it turns out, have actually punished the relevant party for her crime. She may merely have imposed an unjustified harm bearing some resemblance to punishment in its full sense.

This line of argument is unconvincing even if the standard of proof is meant to be our linguistic intuitions. Whether or not we think some party *A* has punished some party *B* in any particular case does not seem to depend on whether or not *A* correctly holds that *B* has actually violated the alleged standard of conduct. A parent may well punish behavior she does not believe to be wrong because, say, her spouse and co-parent takes it to be wrong and she has agreed for the sake of consistency and marital accord to treat it as such. Or, to stay with the uncontested state case, consider the angry, riotous mob employed in objections to a certain kind of flat-footed utilitarian theory of punishment: The mob stands outside the courthouse demanding that Sheriff A should string up Old B for a crime Sheriff A knows Old B did not commit. But Sheriff A

believes with good reason that if he does not string up Old B, the mob will do untold damage, go after Old B's family, etc., etc. If, in such a case, Sheriff A puts on a show of stringing up Old B for the alleged crime, then Old B has been punished for that crime, whatever *A* in his capacity as sheriff may believe. In both cases we have punishment without any actual wrongdoing on *B*'s part, or even the perception of *A*'s belief in *B*'s wrongdoing. These are the very sorts of cases the proponent of the narrow view wants to call at best "sub-standard" or "secondary." Whatever discomfort these cases may elicit, though, it does not seem to be a discomfort about whether or not *A* has punished, but about whether or not *A* has punished permissibly. Indeed, it would be difficult to articulate that (legitimate) discomfort were we *not* entitled to call such impositions punishment, and so unable to level the (warranted) accusation that *A* has punished *B* unjustly.

Even if one were to grant that to sanction the innocent or sanction without proper authority is not strictly speaking punishment, there would still be a further question about condition (v): Why must the relevant form of authority be *institutional*? Why use the definition of punishment to settle in advance that the only legitimate authority to punish is the authority of the state?

Here I have found little by way of clear or explicit arguments, but in its place a general sense that is worth articulating. It is the sense that punishment as it may have existed before the rise or outside the reach of the modern state and its monopoly on violence is something more primitive, and that what the state is up to is, by contrast, something *morally new*. There is something deeply right in this. When the state achieves the vaunted "monopoly on violence," enacting that "violence" according to laws and standards of due process applicable and available to us all, and when that "violence" is enacted, too, on *behalf* of us all rather than for the sake of personal interests or from a base impulse, then a form of response that might once have

amounted to mere vengeance is, indeed, potentially transformed into something else. Even the state execution of a convicted murderer, imposed on purely retributive grounds, is (or at least may be, depending on the kind of state it is) something quite different, which *may* have something more to be said for it than the extra-judicial killing of a murderer by a family member of the victim. In the latter case we have personal grievance answered by personal vengeance. In the former we have the notion that a crime committed against an individual living under the law is a crime against all of us, and so the notion of a debt to *society*. We also have a system in place that aims to make an unbiased determination of guilt. One needn't be convinced that the state is entitled to execute murderers on retributive grounds to hold that this practice nonetheless represents a kind of moral progress.[19]

When Flew and company claim that punishment as it may take place outside of the state context is a merely sub-optimal, secondary, or partial version of what the state does when it punishes the guilty, they seem to be motivated by the desire to rule out some of what they (most of them hybrid theorists) are not interested in defending: either, on the one hand, bare retributivism in the form of unchecked personal vengeance, or, on the other, the flat-footed utilitarianism according to which Sheriff A was right to punish Old B. Hart explicitly denies that this definition is meant to operate in this way—or, rather, chastises those who may have done so.[20] But Hart, like the others, thinks it is sufficient answer to hold these forms of treatment to be

---

[19] I do not mean here to argue that state execution is necessarily or in fact morally better or more defensible than personal acts of vengeance. Indeed, it may be something much worse. I mean only to point out that this position has at least some initial plausibility.

[20] Near the beginning of "Prolegomenon to the Principles of Punishment" when Hart spells out the five conditions, he makes a point of saying that this definition does not, nor is it meant, to license the use of what he calls the "definitional stop." He says explicitly that it "is an abuse of definition" to define punishment so as to rule out in advance those cases that might make trouble down stream for one's preferred justification. He is concerned, in particular, to ensure that this

sub-optimal punishment (rather than no punishment at all), and offers no further reason for thinking that punishment that fails to meet criteria (iii) or (v) really is "sub-optimal" in any conceptual (as opposed to justificatory) sense. His reasons for marking out interpersonal punishment as suboptimal seem no different in kind, nor any more convincing. Flew and company—along with many contemporary philosophers— want to talk about state punishment, treat it as distinct from punishment in whatever other forms it may exist, and, crucially, they want to mark out its replacing of informal systems of punishment as a distinctive human achievement.

Accommodating these considerations, though, does not require us to twist ourselves into these conceptual knots, cluttering up our definition of a morally fraught but relatively simple form of action. We must begin by asking ourselves not what the word "punishment" means, but what punishment is. What, fundamentally, is it to punish? Punishment is the imposition of hard treatment in response to wrongdoing. That is its essential character. There is no deep problem about how to define punishment, or about what belongs to the definition and what belongs to the justification. It is also no degradation of state punishment to understand it as one of several forms that punishment—even legitimate punishment—may take.

The full argument for these conclusions will require several chapters, but for now let me propose a more precise alternative definition of punishment. On my understanding punishment has three necessary and jointly sufficient conditions: That it constitute a deprivation or burden, that it constitute a response to some alleged or perceived wrong, and that it be intentionally imposed. Flew's conditions (iii) and (v) are dropped entirely, and the remaining three conditions

---

definition not be employed by those who prefer a utilitarian justification for punishment to rule out in advance the possibility that a government might punish the innocent for the sake of the greater good (Hart 1960, p5).

are variously weakened and rendered more precise. It is a view on which a much broader swath of human behavior will count as punishing than philosophers have traditionally characterized as such. It is a more principled account of punishment— one on which what unifies the set of behaviors is nothing to do with the regularities of language use across a particular set of language users, but with the structure of the underlying intention. Punishing, on this view, is a particular kind of action, distinct from other kinds of action in virtue of the kind of intention it involves. It is an action that can in principle be undertaken by any kind of agent—by any kind of thing, that is, capable of intentionally imposing some deprivation, and having, as a reason for this imposition, that the object of their intentional action has done something wrong. I will argue that we ought to understand punishment, as a general category, in the following way:

> *A punishes B where: in response to some alleged or perceived wrong on B's part, A imposes some hard treatment upon B.*[21]

This, I will argue, properly characterizes of all of the paradigm cases of punishment, and any other circumstance that it properly characterizes is one that we ought also take to be an instance of punishing. Implicit in this characterization are a set of necessary and sufficient conditions on some form of treatment amounting to an instance of punishing. These conditions are as follows:

- *The Hard Treatment Condition. A*'s treatment of *B* should, at least as far as *A* knows, constitute some burden or deprivation for *B*.
- *The Wrong-Responsiveness Condition. A* should perceive or allege some wrong on the part of *B* to which *A*'s treatment of *B* constitutes a response.
- *The Imposition Condition.* The fact that *A's treatment of B will constitute a burdening or deprivation of B* is among A's reasons for treating B in that way.

[21] I am not here treating "punish" as a success term. My question is about the conditions under which *A* is a punisher, which may or may not come apart from the question of the conditions under which *B* is punished.

Activity that meets these three conditions has punishment's structure and requires punishment's justification.

This understanding of punishment is a capacious one. Consider first condition (1)—the Hard Treatment Condition. The Hard Treatment Condition takes the place of Flew's condition (i)—that punishment be "an evil" or "unpleasantness" to the victim. The Hard Treatment Condition is quite broad. Whether or not it is any broader than (i) will depend on how one understands the notion of an "evil". The Hard Treatment Condition is not a requirement that the punished should be made worse off all things considered. It does not specify that any particular kind or degree of suffering be involved. To meet this condition, *A*'s treatment of *B* need only constitute, at least as far as *A* knows, some sort of deprivation for *B*, regardless of how, or how deeply, the deprivation or burden in question is felt. The Hard Treatment Condition assumes that any desire, aversion, right, or reasonable expectation can, in principle, be leveraged to punish. *A* can punish *B* by giving her what she does not want, or taking from her what she does. *A* can punish *B* by taking from her what she has, or what she has a right to, or what she was expecting. What matters is *that A*'s treatment of *B* constitutes some burden or deprivation, and not *what sort* of burden or deprivation it constitutes. [22]

[22] One might object that the Hard Treatment Condition as I have proposed it is too permissive if it includes forms of treatment that would only count as a mere *nuisance* or even *no trouble at all*. Suppose I were to retaliate for a perceived slight on the part of a friend by taking some small item from his home—a pen or knickknack. Suppose it is an item about which he does not care in the least. Suppose, in fact, that he never even notices its absence. Can I really be said to have punished him, even if all further conditions on punishment are met?

I am inclined to say here that this is an instance merely of ineffective punishment. Consider the case of the person who wants, for reasons of his own, to be imprisoned, who breaks the law as a means of achieving this end, and experiences only relief when he is escorted to his cell. (For an interesting discussion of a set of real-life cases in which people commit crimes for the purpose of being incarcerated, see Dolovich 2012, p1087-1099.) We should probably conclude that in a case like this one, something has gone wrong—that some aim the state has in

There are particular varieties of burdens and deprivations that we strongly associate with punishment. These will tend to include those deprivations standardly or historically deployed by the state or the caregiver of children. It makes sense that such regularities would emerge, given the sorts of regularities there are in human rights and preferences.[23] These sorts of regularities may prime us to skepticism when it comes to claims that some particular deprivation amounts to a punishment if it is not on the list of deprivations typically employed that way in the paradigm domains. In Chapter 3, for example, I will claim that the silent treatment can be a form of punishment, where this involves refusing for a time to speak or otherwise engage with a friend. This, and a range of more subtle form of characteristically interpersonal deprivations and burdens, will meet the Hard Treatment Condition, though they may not be on the list of deprivations and burdens one is used to thinking of in this way. But it is not a requirement, on my definition, that *A*'s treatment of *B* should take the form of one of a list of familiar deprivations or burdens commonly imposed as punishment in the instances with which we are familiar (e.g. incarceration, spanking, the stockade). Any such list is very likely to be relative to a particular culture or set of experiences. Flew and company, who took themselves to be analyzing language use, preferred to think of these changes over time as a sense in which the

---

punishing has likely not been served. It would be incorrect, though, to say that the person so incarcerated was not thereby punished. Being deprived of liberty is sufficient for punishment—unsatisfying or ineffective though it may be—whatever the subjective experience of the deprived. Likewise, in stealing something belonging to a friend, I thereby deprive her of something to which she has a right, and this can constitute a punishment even if it turns out that she doesn't care in the slightest, and even if she experiences the loss as a benefit. What we have is merely a case of (perhaps comically) ineffective punishment.

[23] The state, insofar as it must set general policies, applicable to all, will settle on punishments that at least in principle constitute a deprivation to anyone to whom they apply, and to which as many people as possible are, in practice, averse. Likewise parents, though their efforts are not coordinated in the way of a system of courts, will in many instances converge on a particular form of punishment because of the regularities that exist among children in terms of their aversions and preferences.

definition of punishment might simply change from one context to another. Recall Flew claims that in an earlier era physical suffering was a necessary feature of punishment, though we no longer use the term in that way.[24] I am rejecting that methodology in favor of something more principled. I aim to establish a definition of punishment that picks out a more general form of treatment as it has manifested, does manifest, or might manifest in the future. If we want to determine what punishment is, it may well make sense to begin from a list of things we know to be punishment, but if we abstract the right principles from the list, they should capture forms of treatment that might be used as punishment elsewhere.

The second condition on punishment is the Wrong-Responsiveness condition. The Wrong-Responsiveness Condition approximates (but widens)'s conditions (ii) and (iii)—that it be for an offense and of an offender—appropriately combined and weakened so as to avoid the problems raised by (iii). To meet this condition, *A*'s treatment of *B* has to constitute a response to what *A* alleges or perceives as a wrong on *B*'s part.[25] This is just to say that *A* always punishes *B* *for* something, and that this something will constitute the violation of some norm of correctness. The state punishes the criminal for doing what is forbidden under the law. The parent punishes the child for doing what is forbidden under the rules of that household. Insofar as human beings have conceived of a punishing God, that punishing God is understood to punish sinners for doing

[24] Flew 1954, p293.

[25] While punishment, strictly speaking, must constitute a response to wrongdoing, there is certainly much treatment we call "punishing" which does not. Consider the case raised by Avishai Margalit of the military recruit subject to hard-treatment in the course of basic training (Margalit 1998, p266-267). The drill sergeant may treat the recruits under her command in much the same way a warden might treat an inmate. We may call this treatment "punishing", but it is not a punishment. It is, rather, a form of training associated with a position of relative honor, as opposed to the position of *dishonor* that the prisoner is in. These sorts of deprivations, while they may meet the Hard Treatment and Imposition Conditions, are not punishments because they do not constitute responses to wrongdoing.

what is forbidden under divine law. Across all of the paradigm domains, punishment is always a response to some perceived or alleged wrong, though not in all cases the wrong of law-breaking.

The "alleges or perceives" constitutes the major sense in which the Wrong-Responsiveness Condition constitutes a weakening of Flew and Co.'s condition (iii). It does not require that *B* actually be guilty of the violation of which he stands accused. *A* can mistakenly perceive *B* as having done something he did not in fact do. *A* can also mistakenly perceive something *B* has done to be wrong when in fact it is not. And the Wrong-Responsiveness Condition can be met by way of the other side of the disjunct: *A* can have *alleged* some wrong on the part of *B* without believing *B* is actually guilty, or that what she has done is really wrong. This allows the Wrong-Responsiveness condition to escape the problems raised by (iii), while still capturing the essential connection between punishment wrongdoing. It does not run us into the problem of under-inclusivity, nor pre-judge at the level of definition what ought to be a issue about justification: Could *A* ever be entitled to punish *B* for a crime that *B allegedly* committed, even if *A* himself does not believe she really committed it? A good definition of punishment will help us to better understand what this question asks, and perhaps even help us to see that the answer to the question is a relatively obvious one. What it won't do is render the question itself ill-formed.

So far I have defended two conditions on punishment: The Wrong-Responsiveness Condition and the Hard Treatment Condition. I have argued that without some such requirements on what counts as punishment our conditions would not rule out enough, and I have defended weak versions of both, on the grounds that stricter requirements would rule out too much. These two conditions alone, though, are not enough. They are necessary, but not sufficient to pick out all and only those forms of treatment we call punishment. There are many instances in which the

state or a parent responds to wrongdoing in a way that constitutes a burden or deprivation for the wrongdoer without being an instance of punishment. When the state removes a child from an unsafe home and places her in temporary care, this may well constitute a terrible deprivation for the parent. However profound the deprivation, though, the state has not thereby punished the parent. This is true even when the intervention occurs in response to some particular act of abuse or endangerment on the parent's part. We can generate similar cases in the context of childhood punishment.[26] In either sort of case we have conduct that fails to live up to some applicable standard and a response to it that hurts, but we do not thereby have punishment. Even in the paradigm domains, we need a further condition.

The third and final condition I have offered is the Imposition Condition. The Imposition Condition approximates Flew's condition (iv)—that punishment "be the work of personal agencies"—but in this case makes it both stronger and more particular. It tells us, first of all, that the deprivation or burden imposed in punishing must be not only the work of an intentional agent, but must itself be intended. What this requirement comes to will depend on the details of one's theory of action, but the general idea here is that the deprivation or burden in question must be neither *accidental*, as the bruise inflicted in breaking up the fight, nor *incidental,* as the suffering involved in the removal of a child by the state. It is not sufficient to foresee that one's act will be deprivatory. That it is deprivatory must be (at least part of) the point.

The Imposition Conditions does not require that depriving or burdening should be the *whole* point. It is as a requirement *that among A's reasons* for treating *B* as she does is that (she takes it) treating him in this way will constitute a deprivation or burden for him. It is not a

[26] Such an example might include a parent who intervenes in an altercation between siblings and, in separating them, inadvertently bruises one.

requirement that this be *A*'s *only* reason for acting in this way. It is just that claim that it be *among A*'s reasons, and that *to this extent* the treatment constitutes a punishment. So, for example, the state may incarcerate a violent criminal *both* as a means of depriving her of her liberty *and* as a means of limiting the immediate danger she constitutes to herself and others. It is only punishment, strictly speaking, insofar as the purpose is to deprive.

The Imposition Condition is not a requirement that the punisher relish punishing, or that he take any satisfaction in it at all. It is true that the fact that *A*'s treatment of B will constitute a burden or deprivation is supposed to count in favor, from *A*'s point of view, of treating *B* in that way. *A*, then, takes some kind of pro-attitude toward the prospect of so depriving *B*. However, the condition specifies nothing about what feelings *A* has at the prospect of so depriving *B*. It requires only that she sees the fact of *B*'s being so deprived, by her, as something to be brought about, whether she likes it or not.

Finally, the Imposition Condition does not require any explicit reasoning on the part of the punisher. The requirement, rather, is, first, that *A*'s treatment of *B* be intentional—which is to say, *A* must be acting for reasons. It also requires that *among* her reasons is one in particular *that this form of treatment will constitute a burden or deprivation for B.* There are thorny issues in the neighborhood about how cognitively demanding it is to act for reasons. I take it, though, that acting for the particular kinds of reasons the Imposition Condition requires is no more or less cognitively demanding than intentional action more generally. That is to say, it is something we are capable of doing on the fly, without doing much by way of careful or even explicit reasoning. Intentional action need not be deliberate action, and actions that meet the Imposition Condition needn't be the result of any explicit deliberative process.

This, then, is the overall picture of punishment as a form of treatment: Punishment is something essentially simpler in its nature than what has been represented to us in earlier views. It does not require, as a definitional matter, the apparatus of an institution. It is, rather, a way of treating someone. It is a way of treating someone that has a characteristic effect (that it deprives or burdens), but more importantly—at least for the purposes of assessing what if any permission *A* has to punish—it is a way of treating someone that is constituted by the *intention* to have this effect, and intending it *because* that person has (at least allegedly) committed some wrong. Anything—whether it be an individual or an institution— that can intend to impose a deprivation, and can intend to impose it as a response to wrongdoing, can be an agent of punishment.[27] Any deprivation, imposed for this sort of reason by this sort of thing, will constitute a punishment.

Understood in this way, one should begin to see punishment as familiar part of human life more broadly. We have likely all been treated this way over the course of our lives—by parents, friends, neighbors, partners, and siblings—and we have likely treated others this way in turn. Punishment can, on this understanding, take the form not only of a swat or a timeout, but also of the silent treatment, a cancelled plan, or a cutting remark. Which isn't to say that all instances of the latter will count as punishing. Often enough these forms of treatment will not meet the Imposition Condition, which is to say that while the treatment in question will be hurtful or otherwise deprivatory, that will have in no sense been its point or purpose. I may go

[27] There are no doubt complexities about claiming that institutions are agents and the kinds of things that can act for reasons. I am not, as a general matter, meaning to make any specific claims about the metaphysics of institutions in general or representative democracies in particular. I only mean to be making what I hope is the relatively uncontroversial claim that states (whatever kinds of metaphysical entities they might be), like individual people, *do things*, that they (states) are the kinds of things that are governed by some norms of conduct and character, and that they should and often do offer justifications for the things they do.

silent because I am too angry, still, to speak in an even tone. I may make a cutting remark from indifference to your hardship than for the purposes of causing it. But where such treatment *aims* to have this effect, that treatment has punishment's unique form, and demands its special justification.

That treatment constituting punishment turns out to be a common feature of human life should come as no surprise, whatever attitude of approval or disapproval one may be inclined to take toward it. Punishment is not an idea that began in the context of the modern state—one that we then, occasionally and illegitimately, import into our personal lives. Nor is it an inherently primitive form of treatment that finds its only legitimate/appropriately fastidious form in the context of the (sufficiently) just and efficient modern state. Punishment is a form of response to wrongdoing as old as agency itself—agency, at any rate, that involves some sense of itself as standing in relationships to other "agencies" where those relationships have terms that may be violated.

Conclusion

The question which is not settled in advance by punishment's definition is whether or not we might ever have the authority in our personal relationships to engage in any of the forms of treatment that are punishing. We are not, as individuals in the context of the modern state, entitled to, e.g., kill or incarcerate those who have done us harm. But are we entitled to intentionally impose other kinds of burden in response to wrongdoing, such as those that involve withholding some normal expression of experience of the relationship that both parties value?

Punishment is a form of treatment common to a whole range of human relationships, institutional and otherwise, the permissibility conditions on which are highly sensitive to the

particular kind of relationship inside of which the punishment takes place. What counts as a(n appropriately) triggering norm-violation will depend on what the terms of that relationship are. What form the responsive hard-treatment can permissibly (or even conceivably) take will depend on a variety of factors. It will depend on, e.g., what of value such relationships have to offer for its members, and what if any of the valuable thing one is entitled ever to withhold or impose limits upon. The nature and limits of the authority one has to impose hard treatment as a response to wrongdoing will always depend on the more general authority one has as a member of that particular kind of relationship. The authority one has as a mother or a neighbor or a friend are not best thought of as lesser forms of authority than that of the state, but as very different ones.

Over the course of the next several chapters I will explore the different forms authority to punish might take in the context of several kinds of interpersonal relationships. I will then return to the subject of state punishment, offering a relationship-centered account of its value. I will end by making some initial comments on the nature and place of self-punishment.

Chapter 2. The Participant View: A Theory of Childhood Punishment

> "*Parents and others concerned with the care and upbringing of young children…are dealing with creatures who are potentially and increasingly capable both of holding, and being objects of, the full range of human and moral attitudes, but are not yet truly capable of either. The treatment of such creatures must therefore represent a kind of compromise, constantly shifting in one direction, between objectivity of attitude and developed human attitudes. Rehearsals insensibly modulate toward true performances*." (Strawson 1965, p9)

Moral failure is a natural and inevitable part of human relationships. Certain forms of interpersonal conflict and wrong-reactive attitudes (such as disappointment, resentment, and guilt) are sometimes appropriate reactions to moral failure. Preparing children to be full and decent participants in this aspect of human life requires providing them with some practice at the uncomfortable experiences of being both the object of and subject to these attitudes. It requires that we practice with them the vital rituals of rupture and repair (e.g. rebuke, disengagement, apology) that they will need to employ in the appropriate expression of those attitudes. As I will argue, some of the ways we will treat children as we engage in the necessary rehearsing of interpersonal conflict will inevitably count as punishing. Some practice of punishing will therefore be an indispensable tool for those tasked with raising children to be competent moral agents.

Philosophers have addressed the question of childhood punishment only rarely and for the most part in passing, and so we lack a good theory of the practice. Standard theories of punishment have focused with near total exclusivity on punishment by the state. Such theories often begin from a notion of punishment as a form of retribution or deserved suffering, which is simply out of place when the objects of punishment are young children. Standard theories also begin from a notion of the state as impartial and so legitimate agent of punishment. In the case of childhood punishment, by contrast, the default agent of punishment is the parent or primary

caregiver.[28] While the parent-child relationship is not one entirely untethered from concerns of justice, these are not the concerns at its heart. It is a deeply partial and ultimately personal relationship, by its nature both caring and tutelary. It is only natural that the treatment of children by their parents, punitive or otherwise, will have distinctive aims, conditions on permissibility, and be expressive of a range of emotions and attitudes beyond those appropriately or even conceivably expressed by the state.

Those philosophers who have directly engaged with the subject of childhood punishment have typically assumed a simple account, on which punishment is a form of discipline that leverages aversion to the unpleasant to ensure that children refrain from doing what is forbidden.[29] Psychologists have explicitly defined punishment in these terms.[30] If this is right, though, childhood punishment is vulnerable to two forms of skepticism about its permissibility as a practice. It is vulnerable, first, to charges of being cruelly manipulative. Children are extraordinarily vulnerable to their caregivers, and are owed at least some limited version of the respect owed to mature agents. A practice that aims to manipulate a child's beliefs and behaviors by the imposition of a harm could seem to run afoul of basic duties of care and respect owed especially by parents. The simple account is also vulnerable to evidence purporting to show that punishment turns out to be a relatively ineffective means of discouraging the "problem behaviors" to which it is a response. Such evidence is sometimes cited by psychologists to

[28] I will use these two terms—"parent" and "primary caregiver" —interchangeably. By each term I mean simply that person or set of persons with whom responsibility (moral, not legal) for a child's well-being ultimately rests. The parent in this sense may be a biological grandparent, an adoptive parent, or some larger collective of responsible adults.

[29] See, e.g., Hampton 1984, and Mangasarian 1894.

[30] Alexander 2011.

suggest that the practice of punishing children turns out to be, at best, an understandable (though ultimately self-defeating) expression of parental frustration. [31]

A theory with the resources successfully to defend or even properly to describe the familiar practice of punishing children will have something more to say about the nature and function of childhood punishment than that it leverages aversion for sake of teaching children what is forbidden. I suggest that such a theory will include accounts of five basic features of the practice, the first two being its *boundaries* and its *aims.* An adequate account of boundaries will tell us what does and does not count as an instance of punishing in this domain and why. There is disagreement among philosophers, psychologists, and in our public discourse about whether, e.g., time-outs or even bare expressions of disappointment constitute alternatives to punishment, or simply *alternative forms of punishing*.[32] An adequate theory of childhood punishment will provide a characterization of the practice to which one can appeal in settling questions about what, in principle at least, counts as an instance of it. Having set the boundaries of the practice, a theory of childhood punishment should also be able to tell us what the justificatory aim or value of such a practice might be. It should tell us something about what value we achieve or serve when we punish children, about what of value might be lost were we to abandon that practice, and about why, more generally, authority to punish in this context rests with the partial.

The third feature of punishment is the *method* or mechanism by which it operates: If punishment's value *does* lie in the service of some further end, how do the forms of hard treatment that constitute punishment work to serve the relevant end? An account of *methods* should tell us something about the mechanisms by which punishment works to serve those ends

[31] See, e.g., Khazan 2016.

[32] See e.g. Lewis 2015, Halford 2009.

for the sake of which we punish. A theory of punishment that provides a plausible account of this feature will have something to say in response to the skeptical worry about punishment's general effectiveness.

The fourth feature of childhood punishment of which an adequate theory ought to offer some account is its distinctive form of *authority*. Though permission to punish sometimes rests with secondary figures like teachers or grandparents, a child's primary caregivers are the default agents of punishment. Their authority seems to be relatively exclusive. It is standard across a wide range of childrearing schemes that only those with some larger responsibility for a child's wellbeing may punish. Being yelled at by a stranger to quiet down may in fact quiet a child much more effectively than being treated in this same way by a parent, but this fact in itself does not seem to entitle the stranger to yell. What, then, beyond mere effectiveness, could explain this special permission of the parent as against the stranger? A sufficient theory of childhood punishment should have something to say about why, as a general matter, the parent has and the stranger lacks the right to punish. It should also have something to say about why strangers, grandparents, teachers, and others *do* seem to be entitled to intervene in this way when they are.

Finally, an adequate theory of punishment will explain the *permissibility* of the methods it specifies. It will provide an answer to the question "why are we entitled to serve that end, however valuable, by means of *hard treatment*?" The view should say something about how the ends for the sake of which we punish children could justify hard treatment in particular, especially when imposed on children by those meant to care for them. To explain this feature of the practice is to speak directly to the skeptic who worries that punishment, as a general form of treatment, is objectionably cruel or manipulative, at least in the context of a caring, tutelary relationship.

If we want a theory of childhood punishment with the resources to specify these basic features of the practice, we should reject those that claim it is merely a way of manipulating belief or behavior by leveraging aversion. We should understand it, instead, as a form of guided initiation into a fraught but inescapable and ultimately valuable domain of human life—namely, that of wrongdoing and redress. In what follows I will defend a version of this position, which I call the Participant View for its distinctive emphasis on what Peter Strawson has called the "reactive" or "participant" attitudes. These include, on the one hand, attitudes like resentment, contempt, indignation, disappointment, and, on the other, gratitude, respect, and certain forms of mature love. I follow Strawson in taking these attitudes to characterize personal engagement in relationships between mature moral agents, and propose that we come to understand moral education in general, and punishment in particular, as in part a process of developing in children the skills, capacities, and dispositions required for the appropriate deployment of and responsiveness to such attitudes. The justifying aim of punishment, I argue, is to develop the skills, capacities, and dispositions required for the appropriate deployment of and responsiveness to, in particular, those attitudes specific to wrongdoing and interpersonal conflict.[33] This view will offer a way of understanding at least some forms of punishment imposed for the sake of a further good as consistent with both care and respect. It will also offer a more robust account of the link between the practice and the aims it serves, thus rendering it less vulnerable to claims that that we could more effectively serve that aim by some alternate means.

[33] The view of punishment I advance is this paper begins from the assumption that interpersonal conflict and the difficult emotions and attitudes distinctive of it are sometimes appropriate, and, ultimately, a valuable part of human life. I defend this view more explicitly in chapter 3. I draw especially, here, on the work of Susan Wolf, and, in particular, on her paper "Blame, Italian Style" (2011). What I refer to as the "wrong-reactive attitudes" are something (though not perfectly) like what Wolf calls "the angry attitudes."

Standard Accounts of Punishment and Their Limitations

The punishment of children is not the kind of practice that our standard theories of punishment—ones that focus on the punishment of adults, administered by the state—can be easily extended to explain or defend. These theories have traditionally been organized around a distinction between instrumental theories, on which punishment is justified as the instrument of some further good, and non-instrumental theories, on which punishment is justified as a good in itself. There are a variety of instrumental theories, each positing a distinctive further end for the sake of which we may sometimes justifiably punish (e.g. to teach the wrongdoer a lesson, to protect the citizenry, to affirm the dignity of the victim). Non-instrumental theories tend to be of a single kind or character. These theories of punishment are *retributive*, claiming that to punish permissibly is a special case of giving a person what she deserves, where giving her what she deserves is in itself important.[34]

To see that retributive theories are ill-suited to the purposes of explaining and justifying the punishment of children, we need look no further than the specification they offer of punishment's *aim*. The central claim of such theories is that the good of punishment lies in giving those we punish what they deserve. It is, correspondingly, a condition on punishment that the punished deserve the particular brand of suffering or deprivation in which the punishment consists. Children, though, are simply not the sorts of agents who can *deserve* suffering or

[34] One could hold a non-instrumental view of punishment that is also non-retributive. On Warren Quinn's view, for instance, the *threat* of punishment is justified on grounds of deterrence or self-defense, but the punishment itself is justified as a sort of subsequent promise-keeping (1984). In this way the punishment itself works not in service of some further good, but simply follows through on an earlier commitment. These sorts of theories, though, are relatively few and peripheral.

deprivation. They are, along with the seriously mentally ill, paradigm cases of persons who lack the rational and self-regulative capacities necessary on most any retributive account for attributions of the sort of moral responsibility that can render one even potentially deserving of punishment. The sort of person with whom a theory of childhood punishment is concerned will be precisely those young persons for whom the relevant capacities for moral understanding and self-control are still too underdeveloped to ground an attribution of desert.[35]

Retributive conceptions of punishment also run us into trouble when it comes to explaining the special nature of *authority* in this domain. Even if it did seem plausible that the central aim of punishing children is the cancelling out of some moral debt incurred by child wrongdoers, it would be hard from here to see what deep reason there could be why punishment should take place in the context of a partial relationship—no reason, that is, why authority to punish should be specially vested in the parent. The reasons look at best contingent—i.e. that the parent happens, in many cases, to be the one with the most reliable access to both the child and the facts that bear on the question of what she deserves in any particular instance. Retributive theories fail to explain why the stranger who *does* happen to know that a child deserves punishment and has the means to impose it might nonetheless lack the authority to intervene. Retributive theories alone fail to explain why the authority to punish should be the special province of the partial.

[35] For purposes of this paper I will use "child" to mean a minor who has not yet met the relevant threshold for responsibility with respect to the case (or kind of case) at hand. I follow Tamar Schapiro here in thinking that there is an "intermediate category" into which "many of the people we conventionally call children" will fall. A person falls into this intermediary category when "they have adult status with respect to some domains of discretion, but not others" (Schapiro 1999, p734). So, e.g., the fourteen-year-old may be a child with respect to his ability to resist certain forms of peer pressure, though he is no longer a child with respect to his ability to get himself ready for school in the morning.

Retributive theories also posit a justificatory aim for the practice that is perverse in the context of the care-giving relationship. One desideratum for any theory of punishment should be that the set of aims and motives it picks out as permissible be consistent with the ideals of the particular relationship in which the relevant punishment practice occurs. While retributive aims and motives may or may not be consistent with an ideal of the relationship between citizen and state, they do not find a natural home in an ideal of parenting.

The aim of punishment in the context of the parent-child relationship is not to balance out a moral debt children incur by their wrongdoing; likewise the problem of *permissibility* here is not the problem of showing that children meet any particular standard of desert. It is, rather, a matter of establishing the further good that could justify impositions of hard treatment on those who cannot yet be said to meet any such standard. We should be looking for an instrumental theory of childhood punishment.

If we are looking for an instrumental theory that posits an aim more consistent with ideals of parenting, a plausible candidate is the moral education theory of punishment. On the moral education theory, punishment's justificatory aim is the moral improvement of the wrongdoer herself.[36] Standard moral education theories, though, are not sufficient to the task of providing a satisfying justification of childhood punishment in particular. They begin from a picture of punishment as the imposition of suffering in response to bad behavior, the purpose of which is to teach children that such behavior is forbidden. It serves its purpose by leveraging a child's

[36] This is not to say that such views posit *exclusively* moral-education-type aims. Jean Hampton, for instance, claims that a system of state punishment can and ought to aim at the deterring of crime, but that the state is only entitled to pursue the good of discouraging crime by this particular means insofar as it serve the relevant aim in such a way as to provide the wrongdoer the opportunity for moral improvement (Hampton 1984, p208).

aversion to an imposed consequence to secure some aversion or unwillingness to engage in the objectionable behavior.

Two articulations of the moral education theory operate in the background of most talk about childhood punishment, whether in the popular media, the developmental psychology literature, or in those rare philosophical discussion of the topic. One focuses primarily on behavior, and the other on an underlying set of cognitive capacities.[37] Both get some things right, but are ultimately incomplete.

The first version of a moral education-type theory I will call the *Behavioral View.* It specifies the first three essential features of the practice as follows:

1. *Boundary.* The relevant form of treatment will include any imposition (upon a child) of an "aversive consequence" or removal of a "desired stimulus" made in response to bad or inappropriate behavior;[38]

2. *Aim.* The justificatory aim of such treatment is to discourage the behavior to which it constitutes a response.

3. *Method*. The imposition in question works to serve its justificatory aim primarily by means of disincentivization and aversive association.

Punishment on the behavioral view is the introduction of some painful consequence in response to bad behavior, the memory and future threat of which act to disincentivize or otherwise discourage that kind of behavior going forward—for example, a child sticks her hand in the cookie jar; her parent issues a stinging slap to the hand; she is thereby discouraged by the memory or experience of pain from sticking her hand in the cookie jar again.

As an articulation of the moral education theory, the Behavioral View runs into

[37] In practice, these views are rarely made explicit, and are often blended in varying combinations.

[38] Alexander 2011.

problems immediately with its specification of punishment's aim. If punishment works merely to disincentivize or ensure some aversive association with a particular behavior, then it is not a practice of moral education *per se*, even if our reasons for wanting to discourage the behavior are moral ones. Moral education will at least aim to teach children *that* some behavior or other is wrong, and not simply that if she engages in it, she will suffer. The justificatory aim specified on the behavioral view is not plausibly understood as the end of a moral education practice.

The more significant problem, though, is with the Behavioral View's specification of punishment's *boundary.* The practice, so characterized, is not necessarily a practice of punishing at all. The view begins from a characterization of childhood punishment that is over-inclusive, capturing not only those impositions that we ought to call punishments, but also cases of mere conditioning or behavioral training. On this view, it would count as a scheme of punishment if one were simply to electrify the cookie jar. There is room to debate how exactly one ought to spell out the principle whereby we distinguish punishment from mere conditioning, but it ought at least to exclude those practices with disciplinary mechanisms that seem from the child's point of view to come from nowhere or from nature itself.[39]

These problems with the Behavioral View's specification of punishment's *boundary* and *aim* track the deeper moral worry to which I have already alluded: that perhaps children are owed at least a limited form of the respect and concern for dignity and autonomy we owe to one another as mature agents, and that "punishment" conceived as simply conditioning-by-hard-

[39] Many philosophers, following Feinberg (1965), have held that a set of necessary and sufficient conditions on punishment of any sort must include an expressive condition to the effect that for any form of treatment to constitute a form of punishment, it must constitute an expression of disapproval or condemnation. Morris suggests that it is a logical constraint on punishment that there "be an intention…to convey to the wrongdoer…that the deprivation is imposed because of wrongdoing," which would have the same limiting effect (Morris 1981, p264).

treatment fails to meet even that limited demand. This line of criticism can be at least partially remedied by a second version of the moral education theory that I will call the *Moral Knowledge View*.[40] On this view we aim to teach children good behavior not in a brute way, but by instilling knowledge not only of what is forbidden, but why. On this view the first three essential features of the practice are specified as follows:

1. *Boundary.* The relevant form of treatment includes any intentional imposition of some deprivation or burden upon the child in response to some (perceived or alleged) wrong.

2. *Aim.* The justificatory aim of such treatment is to convey to the punished that her behavior was wrong, in order that she may come to *understand* her behavior as wrong and come, thereby, to experience appropriate remorse. This, in turn, affects behavior going forward and allows for the principled reintegration of the wrongdoer into the moral community.

3. *Method*. The imposition in question works to achieve that aim—of conveying that a certain form of behavior is morally impermissible—by the introduction of some painful stimulus or deprivation where this serves, first, to get the wrongdoer's attention; second, as an expression of disapproval at the bad or inappropriate behavior; and, finally, to either disincentivize or block entirely the possibility of the wrongdoer continuing on as she has, thus providing her with a reason and an opportunity to reconsider the beliefs and attitudes informing the bad behavior.

As the Behavioral View targeted bad behavior, the Moral Knowledge view targets behavior and false moral beliefs such as "I was right to injure him. He had it coming," or, "My interests are more important than the interests of others," and so on. On the Moral Knowledge view, the justifying aim or end of punishment is the opportunity it affords the wrongdoer to correct the false beliefs that lead to bad behavior, and the full restoration to the moral community on which this form of moral understanding is a condition. It shares some specification of mechanisms with the Behavioral View (i.e., the introduction of painful stimuli) but to different, more cognitively

[40] While there is not one consistent view amongst moral education theorists as to how best to characterize punishment's aim and method (views about method, especially, tend to be implicit or cursory), I take this to be a fair generalize, drawing primarily from the moral education theories of Jean Hampton (1984), Frances Gill (2003), and Herbert Morris (1981).

complex ends. It also introduces new mechanisms by which punishment allegedly works to serve its ends: namely, contemplation and the reevaluation of preexisting belief. Punishment, on this view, is meant primarily to make you think.

The Moral Knowledge View expands and improves the Behavioral View along several dimensions. First, it moves away, in its initial specification of the boundary, from a limited sense of punishment as the imposition of suffering, with its connotations of physical pain and psychological abuse, to a broader notion of punishment as consisting in potentially any sort of deprivation or burden, including loss of privileges and the temporary withdrawal of, e.g., the expressions or experience of parental approval or affection. This characterization of childhood punishment as encompassing even rather subtle expressions of disapproval has a long philosophical history. Consider Locke's suggestion that the best way for a parent to punish is simply to "shew a cold and neglectful countenance."[41] Kant similarly recommends the "moral punishment" of children as against physical punishment, where moral punishment consists in "do[ing] something derogatory to the child's longing to be honored or loved."[42] Whether or not one agrees that the imposition of subtle psychological or emotional deprivations is any more effective or less objectionable than the physical or material alternatives, there is good reason to adopt the wider view on which such impositions can and often do constitute forms of punishment.

In Chapter 1 I defended the position that a deprivation needn't be of a particular kind or clear any lower threshold of intensity to constitute punishment. What matters is just that it be *some* form of deprivation, and that it meet certain further conditions (e.g. that it be a response to

[41] Locke 1779, p57.
[42] Kant 1899, p87.

some perceived or alleged wrong on the part of the punished). I see the adoption of this more capacious view of punishment as necessary to a proper defense of the practice. Much of the treatment I defend as permissible punishment across all interpersonal contexts would not count as punishing on a more restricted definition. I also see the more capacious view as necessary for punishment's proper condemnation. If we adopt a theory of punishment on which, e.g., the parent who responds to bad behavior with abusive or manipulative withholding of affection could not in principle count as punishing, we adopt a theory without the resources accurately to capture what is (or at least may be) the problem: That the parent in question is punishing his child, and that he is doing so in a way we should find objectionable.

The Moral Knowledge View also differs from the Behavioral View in its conception of punishment's educatory aims. On the Moral Knowledge View, punishment's target is not behavior, but a set of underlying beliefs, the transformation of which will trigger, in turn, the proper remorse that makes possible the further good of reconciliation. This expansion of the characterization of punishment's aims is a good one, but having made it, the Moral Knowledge View has a harder time offering an adequate account of the method and mechanisms whereby these more expansive aims are achieved. On the one hand, it is not clear that the methods the Moral Knowledge theory suggests are sufficient to serve the aims they specify. Sufficient methods would be those that ensure, first, that children reliably learn by those methods what acts are morally prohibited and why *and* that this coming to understand should suffice to trigger remorse and alter future behavior for the better. It is not clear, though, that the methods spelled out by the moral knowledge theorist are sufficient for either. To believe that the way punishment works is primarily by inviting the punished to reevaluate her underlying moral beliefs is to have an implausibly intellectual view of how punishment works to morally educate in the case of

young children. The moral knowledge theorist holds that our aim in punishing is to bring it about not only that children refrain from doing what is forbidden, but that they act in doing so from a deeper understanding of the moral reasons they have. But there is something implausible in the thought that simply making a child understand what acts are forbidden and why would secure this end. A person who understands what sorts of behaviors are prohibited and why, even one who tends to feel remorse at violating those prohibitions, is not necessarily a person capable of or inclined to abide by them. More is required of an account of methods if we are to accept that the practice actually works to serve its justificatory aim.

That the methods specified on the Moral Knowledge view are not sufficient for achieving the specified aim is not in itself an insurmountable problem, but they also appear to be potentially *unnecessary*. There are a range of experts willing to attest that there are other means more effective than punishment of serving even the basic behavioral outcomes that allegedly justify it.[43] These experts suggest that most problem behaviors are more effectively solved by imaginative exercise and positive reinforcement. If the aim is to understand what acts are forbidden and why, we might simply wait until the child in question is old enough to comprehend the truth about how a person ought to behave and why, and then explain it to her. We can allow her to take risks and to learn from her mistakes. We can encourage her to ask questions, keep a journal of how she sees or puts to work in her life each new piece of moral knowledge as she acquires it through explanation and experience. Thus, it is not obvious that punishment, so characterized, plays an ineliminable role in a program of moral education. We do not (or no longer, at least) take these kinds of inflictions and deprivations to be a crucial means

[43] See, for instances, Karson 2014, Khazan 2016.

of teaching a person to read, or do math, or fly a plane. Why think they are necessary for teaching children what is forbidden and why?

There is clearly something right in these standard articulations of punishment's educatory function, but there is also something missing. They leave us in the position of trying to defend an instrumental theory of punishment on which punishment may turn out to be neither necessary nor sufficient for its allegedly justifying aims, thus leaving itself vulnerable to skepticism about its efficacy. The Moral Knowledge View is still a view on which punishment works by leveraging desire and aversion to manipulate the punished—in this case taking belief rather than only behavior as its target—thus leaving itself vulnerable to the charge that it fundamentally violates duties of care and respect. While the Moral Knowledge View moves us in the right direction, it still falls short. It, too, begins from an impoverished picture of punishment's educational function and aims, and an alternately unconvincing and unsettling account of its methods.

The Participant View

A better moral education theory of punishment will require a better theory of moral education. The picture of moral education through punishment offered on the Behavioral and Moral Knowledge views begins from an understanding of punishment, and of moral education in general, which takes what is for all intents and purposes the child's-eye view of the matter. Understood from this perspective, punishment is a discrete event, the nature of which is that something you don't like happens to you, and the point of which is to teach you a more or less explicit lesson: *Don't do the thing you were doing*. But the tutelary aims of punishment and the mechanisms by which it works to serve those aims are much broader and more complex. We should instead conceive of childhood punishment as an overall practice or pattern of intervention

that works over time to initiate children into a particularly fraught and crucial set of personal attitudes and practices: namely, those appropriately responsive to wrongdoing and the interpersonal ruptures that wrongdoing can constitute.

Moral failure is a natural and pervasive part of even the best human life. All human beings sometimes fall short of any reasonable ideal of friendship, citizenship, neighborliness or collegiality, and so each of us will sometimes wrong others and sometimes be wronged. What comes much less naturally is a mature and decent response to such failure. It is a difficult thing either skillfully to make or to field an accusation of betrayal or disrespect. Yet engaging decently in any mature human relationship requires that we develop this skill. Thus, I argue, the justificatory aim of a defensible practice of childhood punishment is: to teach children to be appropriately sensitive and responsive to wrongdoing and charges of wrongdoing. Learning what behaviors are forbidden is one small part of the project.

Consider a third sort of moral education theory—the Participant View. The Participant View begins from the following basic specifications of the practice, its aims, and methods:

1. *Boundary.* The relevant form of treatment includes any part of the practice of suspending and resuming normal relations with children in response to some alleged or perceived wrong, where the suspending constitutes the intentional imposition of a burden or deprivation as such.

2. *Aim.* The justificatory aim of this practice is to train children to be properly sensitive and responsive to alleged and perceived wrongdoing, cultivating in them the full range of cognitive, affective, and behavioral dispositions necessary for competent participation in the forms of conflict that responsible participation in mature relationships will require.

3. *Methods.* The methods of achieving this aim are myriad, but all operate for the most part iteratively. I focus here on two that I take to be particularly prominent and under-discussed:
    a. *Modeling.* In punishing fairly and proportionally and by expressing apt wrong-reactive attitudes, caregivers model for the child proper reactions to wrongdoing.

b. *Practice.* In being punished, children practice receiving moral criticism of a special kind, thereby learning to regulate their own responses to interpersonal conflict.

1. Boundary.

An adequate theory of childhood punishment should begin from a different account of what is and is not part of that practice. The version I offer here, in the Participant View, starts from the broad understanding of punishment defended in chapter 1, according to which punishment is any intentional deprivation or burden that constitutes a response to wrongdoing. The adoption by the Moral Knowledge View of a more inclusive standard for what might count as punishing—one that includes non-material deprivations— already moves us in this direction. I have argued that what makes some deprivation a punishment is not that it is a deprivation of a particular sort or intensity, but that it is imposed as such in response to some perceived or alleged wrong. Our conception of childhood punishment should include any such deprivation imposed by a mature agent on a minor whose capacities for self-regulation and moral understanding are not yet sufficiently developed to support a reasonable standard of desert (i.e. a child in the sense relevant to our inquiry). Even a parent's aloofness may on this view constitute a punishment if it constitutes a deprivation intended as such.

There is more to be said, though, about the character of the practice whose permissibility we set out to assess, and the nature of the boundary carving out what does and does not count as being a part of it. The Participant View makes three further revisions. First, while earlier accounts sought to justify some particular instance of punishing for its contributions to some further good or value, the Participant View picks out the overall practice or *pattern of interventions over time* as the object of justification. While the punishment of an adult by the state may be characteristically episodic, childhood punishment is not. This is not to say that

episodic missteps do not matter, nor that episodic interventions have no impact. If a parent responds to some particular instance of bad behavior with a sufficiently harsh and startling punishment, this alone will sometimes be enough to discourage a child from behaving that way again, thus achieving one kind of educatory aim one may have in punishing. Insofar, though, as punishment works to serve its justificatory aim of educating by consistent repetition, we had better understand the object of justification not at the level of individual interventions, but the level of the larger pattern of intervention.

The Participant View incorporates into our sense of the practice not only the attitudes and expressive rituals of *rupture*, which will include the imposed deprivation in which punishment is generally thought to consist, but also the attitudes and rituals of *return*—the relief from deprivation; the restoration of the strained relationship. Such relief or restoration will in some cases go unmarked, with things quietly returning to normal, and in others will include more ritualized returns in which, e.g., a child meets the specified condition of apologizing to the sibling she hit, apologies are accepted, and she is ceremoniously welcomed back into the game. Childhood punishment *as* a practice involves more than the intervention itself, but instead begins with the triggering event and ends in whatever implicit or explicit rituals of forgiveness or reintegration signal a return to normal life. Moral education views focused on state punishment have treated the possibility of such reintegration as a further good that punishment serves. The Participant View names reintegration as part of the practice in need of a proper defense rather than a further aim of that practice. It is a distinctive feature of childhood punishment that where there is no subsequent forgiveness and reintegration, the imposition has not merely failed to bring about some further good. Rather, the arc of the intervention is itself incomplete. Failing to

properly resolve the tension that punishment introduces into the relationship is not just a failure of parenting, but a failure of the parent to punish appropriately.

Finally, the Participant View characterizes this pattern of deprivation and its eventual alleviation in terms of "the suspending and resuming of normal relations." This language does not constitute a further revision to the Moral Knowledge View so much as an elaboration of it. Such views already demanded a widening of the picture of what sorts of deprivations might count as punishment to include things like the withholding of praise, and subtle shifts in apparent attitude or emotional valence. What is left is to consider how to characterize this wider category of deprivations in terms that illuminate what unifies them. The initial impulse here, as in the context of any theory of punishment, might be to say that the fundamental commonality amongst the relevant instances of depriving is that they cause the object to suffer, or aim to. But while an instance of spanking might seem to be best characterized in this way, it seems much less helpful as a characterization of, say, a time-out, which *may* involve suffering but isn't necessarily *about* suffering. What spanking and time-out have in common is that a child has acted in such a way that the parent judges that *things cannot go on as they had before.* There is a pause in normal life that takes the form of a deprivation, intended as such. The Participant View thus characterizes punishment in terms that capture the whole range of impositions that we should think of as being part of that practice, and that captures something about what actually unifies them as class.

2. Aim.

On the Participant View, as on the Behavioral and Moral Knowledge Views, the justifying good that punishment of children serves is their moral education. Punishment is a practice justified insofar as it contributes to a developing agent's moral understanding as expressed in (among other things) her treatment of others. It thus contributes to her own

wellbeing as a developing moral agent, and the wellbeing of those on whose lives her actions and attitudes have some impact. Where the Participant View departs from other moral education views is in its assessment of that in which this moral education consists. On the Participant View, punishment works to develop the whole range of skills, capacities, and dispositions required for a just and decent responsiveness to wrongdoing. These skills, capacities, and dispositions will be behavioral, cognitive, and also affective. They will involve not only an understanding of what behaviors are forbidden, but the capacity to recognize such behavior when it happens, and, crucially, the set of dispositions, capacities, and skills required to respond appropriately to wrongdoing when and as it occurs—including those that allow us to regulate the sorts of psychological and emotional reactions that often arise in such contexts.

Adopting this view of childhood punishment's educative aims will, as we shall see, help provide a more satisfying justification of the practice. The first reason we have to adopt the expanded view, though, is not for the sake of ensuring justification, but for the sake of beginning from a full and accurate picture of how the practice does in fact operate. Consider, for example: A child hits her sister, is placed in time out, and is told she can rejoin the group when she has cooled off and is prepared to apologize. The only explicitly stated lesson may be "Don't hit your sister," but this is not the only or even necessarily the most important dimension of the educative function of this intervention. Other relevant, though perhaps less explicit, potential lessons include:

- When you are very upset, you should take a minute to cool off.
- When you have wronged someone, you owe him an apology.
- When you hit someone, it disrupts the relationship and its associated benefits.
- It is consistent with caring about someone to draw boundaries about the kinds of behavior you are willing to tolerate.
- Hitting is among the forms of behavior one should not to tolerate, no matter the provocation.

The lessons punishment teaches—even conceived of as a set of moral propositions about what is true or how to behave—are more than those attributed to it on the Moral Knowledge View. And punishment, on the Participant View, works to teach something more or other than a set of moral propositions, but involves developing a set of skills, capacities, and dispositions.[44]

This revision to the view of punishment's educative aims should be understood in the context of a more global view of that in which a full moral education consists. We should understand a moral education, conceived of in the most general terms, as consisting in the cultivation of proper sensitivity and responsiveness to the morally salient features of the world children will encounter.[45] We aim to prepare children to live—to equip them with the inner-resources they will need to live *well*, rising to the challenges of life as it comes. These challenges will not typically come in the form of an exam, where knowing the correct answer, in the sense of being able to recite the appropriate responses under controlled conditions, will suffice. We do not aim to equip children with a mental checklist of what is good, valuable, required, permitted,

---

[44] There is disagreement among epistemologists on the question of whether the sorts of skills and abilities I have in mind here (referred to in the literature as "knowledge-how") really just amount to the knowledge of proposition (or "knowledge-that"). While I am inclined to follow philosophers like Ryle and, more recently in thinking that know-how is not merely knowledge of a set of propositions (see also, more recently, Glick 2012), the Participant View does not depend on the success of the anti-intellectualist view of know-how. If acquiring the set of skills I have in mind turns out to be a matter of acquiring knowledge of a set of propositions, these will still turn out to be a different set of propositions than those picked out by the Moral Knowledge View.

[45] In moving away from the Behavioral and Moral Knowledge Views, one need not think that there is nothing right in them—only that they emphasize one aspect of moral education's ends to the exclusion of others. This sort of criticism, and the presumed background view of moral education in general, is echoed by the work of psychologists working on questions of moral development. In "Moral Exemplarity" Lawrence J. Walker criticizes the "artificial trichotomy" represented by "the three major competing traditions in moral psychology"—one emphasizing behavior, one cognitive development, and one moral emotions. Such distinctions, he argues, "obfuscate the interdependent nature of thought, emotion, and behavior in moral functioning and trivializes our understanding by an exclusive focus on some particular component that has been hived off" (Walker 2002, p67).

etc., or even merely the capacities requires for making these determinations on their own. Living well amounts, more generally, to being *appropriately responsive* to the circumstances one encounters, where this involves both *recognition* and *reaction*. We aim to teach our children to recognize danger where there is danger, injustice where there is injustice, a good friend in a good friend, and to respond to these *well*—appreciating true friendship, fighting injustice, fleeing or standing to face danger as circumstance demands. Providing adequate training for life also requires some means of teaching children to recognize the relevant features of their own inner-lives, and to respond appropriately here, too—learning to recognize and calm panic as it rises in the throat, to attend to happiness when and as it comes. Real competence, and ultimately mastery, require the capacity to sense and react appropriately to these features of one's inner and outer life across the whole range of conditions one may face, from the calm to the very turbulent.

One way of framing this view of moral education would be to say that the moral competence and ultimately mastery that moral education in general and punishment in particular aims to teach consists in what Ryle called "*knowledge how*."[46] To be properly morally educated according to Ryle's framework is neither simply to behave in a way that happens to accord with a set of moral rules, nor to accept as true some set of propositions about what the moral rules are, nor both together. Moral know-how is, rather, "[a form of] knowledge…actualized or exercised in what [the agent] does."[47] Part of what a child learns by punishment is what the pilot learns by crash simulation, or the police officer learns by emergency drills: to navigate a difficult circumstance ably, despite these difficult and judgment-clouding circumstances.[48] The

[46] Ryle 1945.

[47] *Ibid*. p9.

[48] The Participant View does not assume or depend on anti-intellectualism about moral knowledge. See footnote 44.

developing moral agent learns in being the object of a punishment practice *how* appropriately to express and respond to a particularly challenging set of the participant attitudes.

The particular set of skills, capacities, and dispositions punishment works to train are those one exercises in being both *subject to* and the *object of* these attitudes, the "wrong-reactive attitudes." The notion of a reactive or "participant" attitude is given to us by Peter Strawson. They are those attitudes

> "which belong to involvement or participation with others in inter-personal human relationships…includ[ing] resentment, gratitude, forgiveness, anger, or the sort of love which two adults can sometimes be said to feel reciprocally, for each other."[49]

Strawson distinguishes these from the "objective attitudes" that one might take up toward a patient one is treating or the citizens of a culture one is observing for anthropological purposes. The participant attitudes, by contrast, only make sense in the context of communities and relationships in which we are in some sense *participants,* and in which all members are capable of participating at a certain level. These are the relationships, Strawson argues, in which it makes sense to hold one another accountable. On a Strawsonian account of moral education, we learn by practice how to manage and when to express the full range of participant attitudes, including not only resentment and disappointment, but gratitude, respect, and mature love. Punishment, on the Participant View, is an essential piece of this training, whose distinctive domain is that of the wrong-reactive attitudes.

What we want from a moral education in general, and from punishment in particular, is a form of mastery consisting in a skillfulness that persists in the face of novelty and turbulence. The police officer who makes panicked mistakes in the face of the normal course of her challenging work is not yet a competent police officer, however well she has performed on her

[49] Strawson 1962, p7.

examinations. Learning both to distinguish what is truly fearsome from what merely seems to be, and how to manage fear is part of what competence and eventually mastery of policing skills involve. It is this form of practical mastery, and not merely knowledge of what is forbidden, that we get from punishment when it goes well. When punishment does *not* go well, its primary impact may not be to undermine a child's understanding of what constitutes bad behavior, but a tendency to over- or underreact to bad behavior in both themselves and in others.

3. Method.

Once we have come to understand the punishment of children as a program that aims to develop a set of practical skills, there are some very general things we can say about it as such. To do anything skillfully—whether it be piloting a plane or blowing glass, police work or making a soufflé—it is not enough to have memorized the manual.[50] Mastery of a complex skill requires more. Often we acquire such mastery by a kind of apprenticeship wherein the relevant skill is modeled for us, and we, in turn, have the chance to practice under special conditions. These conditions will include things like real-time feedback and guidance, a slow escalation in complexity or difficulty of the task over time, and a limited (perhaps de-escalating) insulation from the worst consequences of failure (as with training-wheels on a bicycle). There are certainly instances in which punishing a child works to discourage a behavior by leveraging desire or aversion, and instances in which punishment works to improve moral understanding by spurring reflection. Much of what punishment teaches, though, it teaches by the familiar tutelary

[50] "We can imagine" says Ryle, "a clever player generously imparting to his stupid opponent so many rules, tactical maxims, 'wrinkles,' etc., that he could think of no more to tell him; his opponent might accept and memorize all of them, and be able and ready to recite them correctly on demand. Yet he might still play chess stupidly, that is, be unable intelligently to apply the maxims, etc." (Ryle 1946, p5).

mechanisms associated with apprenticeship—namely, modeling and practice.[51] This will be particularly true in the case of young children, with regard to whom the Moral Knowledge theorist's epiphany-spurred-by-somber-reflection notion of how punishment educates looks particularly implausible.

Where an instance of punishment works to serve a specified aim, it does so by providing children with the opportunity to observe and practice certain appropriate forms of interpersonal conflict. Punishment operates by serving as an opportunity for children to observe their parents' reactions to their own bad behavior, with parents serving as a model of what is appropriate.[52] While modeling is a form of discipline that critics often contrast with punishment, punishment is itself a form of modeling. Children do not only suffer their punishments, but observe them being administered. They learn in being subject to a practice of punishment not only what acts are wrong, but what counts as an appropriate reaction to a particular kind of wrongdoing, and how

[51] Susan Dwyer has cast serious doubt on the notion that habituation and training are the primary mechanisms by which developing agents come to achieve moral competence in any given domain (Dwyer 1999, 2003). She notes that "[e]xplicit parental instruction radically underdetermines the child's ability to distinguish between transgressions of different kinds, and [that] children make these distinctions long before they are able to articulate moral (or conventional) rules as such rules" (Dwyer 2003, p189). I agree with Dwyer here. Indeed, my arguments are motivated in part by the conviction that traditional moral education theories have been insufficiently sensitive to precisely this point. Dwyer goes on to emphasize, instead, the role of "innate moral endowments," noting that even very young children are able to, on the one hand, understand and empathize with other persons, and, on the other, discern and even distinguish between moral and conventional rules. Here, too, Dwyer and I are largely in agreement. I do not mean to deny the crucial role that innate capacity plays in moral competence. My claim is only that innate moral endowments, *too*—even in conjuncture with explicit parental instruction—underdetermine the overall set of abilities, capacities, and dispositions required for moral competence. I presume only that proper development of the innate cognitive and affective capacities necessary for moral competence requires a process of socialization and moral education in which explicit instruction plays but a part. Modeling and practice, I argue, also have a crucial (though not exhaustive) role to play.

[52] The idea that exemplars play a crucial role in the development of character and understanding is one we find throughout both the history of philosophy—most notably in Aristotle, but also in Kant (see, e.g. 1781, p134 and 1787, p174)—and contemporary cognitive science (Clark 2000).

they ought to expect others to react, where this will include, especially, the expression of wrong-reactive attitudes like resentment and disappointment. Part of punishment's terrible power, where it goes wrong, is to teach children to express resentment, anger, and frustration inappropriately, and to expect and tolerate such inappropriate expressions from others. When punishment goes well, by contrast, children begin to learn by observation not just where the line is, but also when and how to draw it.

To see how a practice of punishing might work to teach children, by modeling, what the mature experience and expression of such attitudes might look like, it will help to consider an analogy to what psychologists call "child directed speech," and we, colloquially, call "baby talk." This way we have of talking to babies—of exaggerating vowels, syllables, intonation, and facial expression—plays a crucial role in language acquisition, both holding babies' attention and helping them in coming to understand.[53] I am claiming that we do something similar in permissibly punishing. Our reactions may be amplified, slowed, or symbolic versions of appropriate reactions to wrongdoing, which specially focus attention and help children to get the basic grammar of conflict and repair. In grounding a child, or putting her in time out, we engage in a ritualized sending away followed by a conditional reintegration. What the child learns by this is not (if things go well) to put their friends and spouses in time out, but to be capable and willing to, e.g., take and demand certain kinds of space, to self-regulate difficult feelings, to apologize. We exaggerate the registering of the wrong to get children on to when there is a wrong that requires registering, what, very roughly, it looks like to register that wrong, and then how and when to let it go. As with child directed speech, there is a kind of theater involved, but

[53] Matychuk 2004.

not one that requires conscious intent. We naturally talked to babies this way long before we understood its role in language acquisition.

At least as important as the opportunity punishment provides to learn by observation is the opportunity it provides to learn by practice. There is nothing more natural than moral failure and our tumultuous responses to such failures both in others and ourselves. There is little less natural than learning to express these reactions appropriately in real-time. What is natural is to panic. What is natural is the bare instinct to flee or to fight. Acquiring confidence and competence in being the object of disappointment, anger, or even bare judgments of wrongdoing requires practice, much like piloting a plane in a storm or doing police work where tensions run high. Like these kinds of competencies, they begin in non-ideal circumstances—circumstances in which something has already gone wrong, and in which the object of judgment or criticism is typically subject to a range of strong desires, fears, and anxieties. These might include the desires that motivated the original (allegedly) bad behavior, or hard feelings at being thwarted. It may also include deep fears of abandonment or rejection. In any case, they are circumstances in which both inner and outer landscape will likely include obstacles to good judgment and good nature both. Learning how to operate wisely and effectively under these conditions requires first-hand experience—e.g. learning what this kind of anger and anxiety feel like, how and when they usefully inform thinking and behavior, and how to calm them when they do not. These are not skills we learn by testimony or observation alone. They are skills we learn by experience, and by failure. In making children subject to a range of natural and often difficult human responses to bad behavior, we begin to bring them into an important and difficult kind of conversation and help them learn both what it is to participate in such a conversation decently and how to—"[r]ehearsals," as Strawson says, "insensibly modulat[ing] toward true performances."

4. Authority

Next, we come to the question of the parent's special authority in punishing. On the alternative theories we have discussed, it was hard to see what reason there could be for this, other than the fact that parents are typically best situated, in terms of sheer physical and psychic proximity, to assess when punishment is warranted and to impose it to greatest effect. Yet the parent who happens not to be so situated seems to retain this special authority. And there are instances in which a stranger looks better situated to effectively influence a child's behavior or beliefs about what is forbidden by punishment, and seems nonetheless to lack the authority to intervene. Nothing internal to the other theories of childhood punishment we have discusses could serve to explain why this should be.

The Participant View gives us something more to say here. First, the richer account it offers of punishment's character and educatory aims makes it much less plausible that the stranger will be properly situated to serve them. On the Participant View, punishment works primarily by targeting an underlying set of skills, capacities, and dispositions (cognitive, affective, and behavioral). While the stranger may have a perfectly good view of a child's behavior in any given moment, he will be less well situated to understand how any given behavior does or does not reflect these underlying feature of a developing psyche. Two instances of public temper tantrums may be identical from the point of view of the stranger, but reflect, in one case, an underdeveloped capacity for certain kinds of self-regulation and moral understanding and, in the other, that the child in question has missed a nap or is running a fever. Given that we aren't in the business of balancing the scales of justice or conditioning behavior, much more detail about the routines and capacities of the particular child will bear on whether and how punishment is appropriate.

More fundamentally, though, the Participant View is an account on which punishment relies for its efficacy on a continuing, pre-existing relationship between punisher and punished. The view of punishment itself here is framed in terms of the impairment and subsequent mending of "normal relations." While there is something like "normal relations" between child and stranger, it is comparatively thin. There is no background relationship of affection, support, and care to be impaired or restored. The stranger can impose harm or inspire fear, but the mechanisms available to her for doing so are more blunt. Much of what I have suggested might count as being among the more defensible forms of impairment amount to exactly the sorts of subtle shifts in attitude that can only take place against the backdrop of affection, support, approval, and so forth, and which can only be defensibly imposed by a party with the resources and obligation to *resolve* the impairment once introduced.

This last point is especially worth dwelling on for a moment: It is typically parents, not strangers, who are properly positioned to engage in the repair that follows rupture, and it is parents who are responsible for doing so. Modeling what reintroduction into the community looks like is not the kind of thing to which the stranger is typically attentive. This is natural enough, given that they bear no primary responsibility for doing so, and are only capable of it in very limited, individual ways—e.g. accepting apologies and expressing forgiveness. The parent, in contrast, is in a position to, for instance, have a more intimate exchange after a stretch of time has elapsed. This might be a conversation in which the child can feel comfortable admitting fault, discussing what happened, and drawing conclusions together. The parent has, in addition to a relationship of intimate trust, a wider range of shared references and experiences to draw on in having such a conversation. The parent may also in a position to broker repair in other relationships, helping the child to return to a group or activity that her bad behavior disrupted.

And, finally, the parent also has a wider range of resources available for drawing the period of" hard-treatment" to a close—resources that include the possessions of objects or knowledge of activities the child enjoys, and, crucially, the parent's own capacity for the expression of affection and approval.

The Participant View certainly provides us with greater resources for making the case that the parent is better situated for serving punishment's aims. But the deeper explanation it offers for the parent's unique authority to punish is that punishment is, uniquely, the parent's responsibility. A proper moral education is essential to a person's well-being; properly training the affective, cognitive, and behavioral reactions to wrongdoing is an essential part of a moral education; a certain type of childhood punishment practice is, in virtually all cases, indispensable to the project of making children properly responsive to wrongdoing; thus, those responsible for seeing to a child's wellbeing are the ones responsible for ensuring that children are appropriately subject to this practice. It is because a child's primary caregivers are responsible for ensuring that she is appropriately subject to this practice that they are entitled to punish, and to delegate that authority to others where deemed appropriate or necessary. Parents are specially entitled to punish because punishment is the unique solution to a problem that arises specially for the parent. Where those *other* that the parent are entitled to punish, it will be in virtue of the responsibility they themselves bear for a child's well-being and education. The teacher, the grandparent, the babysitter, even the stranger (insofar as they are another responsible adult in the vicinity) may sometimes be responsible in some limited way—typically licensed by the parent—for the child's well-being and moral understanding, and so it makes sense that the permission to punish, too, can travel in certain limited ways, alongside that more general responsibility.

5. Permissibility

We are left now with the permissibility question. What, first of all, entitles us to punish children though they do not *deserve* to be punished? If it wasn't clear before that desert is not the right kind of condition on punishment's permissibility in the case of children, the preceding discussion of punishment's educative aim should help to clarify. To ask whether or not a child *deserves* punishment is tantamount to asking if she deserves to be made to eat her vegetables or to graduate to the fourth grade (though she is afraid of going to the new, larger school, or to face more difficult homework). Whether one should or should not make a child do these things is not a matter of desert but of her capacities and the conditions of her wellbeing.

What entitles us, though, to punish in the service of our aim of morally educating them, and what are the limits of that entitlement? I have argued that some practice of punishing is required of caregivers insofar as it is a necessary means of making children into competent moral agents. I have claimed that some such practice is typically necessary in this way. But what if any limits are there on how we may punish children and when? The Participant View of punishment is a variety of moral education theory and as such is not a deterrence view, but it is a broadly instrumental one. It is a view on which we punish for the sake of a further good—that of seeing to the moral development of those whose moral development is our responsibility. A traditional problem of instrumental views has been to explain why we should not be entitled, e.g., to punish shoplifters more harshly than murderers, if it turns out that shoplifters are more difficult to deter. Such accounts sometimes have difficulty generating the resources necessary to explain fairness or commensurability as conditions on punishment when such considerations come apart from effective deterrence. The Participant View might seem to have a similar problem: In adopting such a view, are we committed to the notion that whatever is necessary to serving the aim of moral education is permitted, however seemingly unfair, manipulative, or cruel?

There are a couple of things that the Participant theorist in particular can say here. First, insofar as punishment teaches appropriate moral responses by modeling and practice, there is a necessary link between a punishment's being permissible, and a punishment's amounting to the kind of treatment that looks to be fair and respectful, thereby effectively *teaching* fairness and respect. For some particular practice of punishing to be permissible on the Participant View, it must constitute an effective means of teaching children how to live well. Inasmuch as a child learns from her parents' treatment both how to treat others and what types of treatment to tolerate from others, her parents must treat her in ways that model these appropriately, even in punishing. If yelling at others is a behavior that is only very occasionally appropriate, probably one ought not model it as a consistent response to the bad behavior of one's child. If hitting others is never the appropriate response to bad behavior, probably one ought never model it as a response to the bad behavior of one's child. Because of the moral nature of the lesson and the modeling dimension of how that lesson is taught, absolute moral constraints on how we are entitled to treat mature agents not in our care will typically have some strong (though defeasible) correlative constraint on how we are entitled to treat children in punishing them. While the Participant View permits hard treatment for the sake of a further good, there are constraints built into the theory that rule out, in most cases at least, those forms of hard treatment that are otherwise unfair, disrespectful, or cruel.

The deeper point, though, is that punishment on the Participant View does not necessarily or even typically work simply by leveraging desire and aversion to manipulate behavior, belief, or anything else. The hardness of "hard treatment," done well, has its source in the difficult nature of the attitudes it expresses, and in the sense of loss and fear that even the appropriate expression of these attitudes involves. It can be startling to experience anger, and frightening to

be the object of it. But we do not merely use that discomfort as a means of manipulating behavior, or of working the ground to loosen children up for receiving an important message about what is forbidden. We are, rather, practicing it. We are practicing together how to be scared, how to feel vulnerable or frustrated, which are not, after all, experiences one could avoid, nor are they ones we should hope to.

Conclusion

The Participant View of childhood punishment is in some ways quite sweeping in its revision of how punishment morally educates, but it begins from three relatively modest thoughts: That interpersonal conflict and the wrong-reactive attitudes (disappointment, resentment, etc.) are a pervasive, inevitable, and important part of the life for which we as caregivers are tasked with preparing the children in our care; that to prepare children for this aspect of life requires that they have ample opportunity to observe and participate in the relevant rituals of rupture and repair by which those attitudes are appropriately expressed; and that some of the ways we must treat children as we engage in the necessary rehearsing of scripts and rituals of interpersonal conflict will be punishing.

None of these claims are uncontroversial, but for those inclined to think they are basically correct, the Participant View offers a better way of understanding punishment in this domain. When we expand our conception of the practice of childhood punishment and its aims, making each more principled and complete, we are able to move away from a picture on which the connection between punishment and moral education may look dubious or insecure. When we conceive of the practice and its aims more narrowly, the legitimacy of the practice may look vulnerable to being undermined by simple empirical investigation or appeals to the rights of

children to basic forms of safety and respect. But if the Participant View is correct, then while the effectiveness or permissibility of particular *sorts* of punishment may be undermined in this way, the practice as a whole will not be. Without it, we would have no way of preparing children competently to navigate a part of human life that, while sometimes fraught and often unpleasant, one must learn to navigate if one is to be a responsible citizen of any human relationship. In being subject to such a practice we learn, where things go well, to be people neither conflict-prone nor conflict-avoidant, but prepared, when necessary, to address wrongdoing forthrightly, and once resolved, to let it go.

Chapter 3. Between Friends: Punishment in personal relationships of equality

I have argued that the value of punishment in childhood is as a kind of rehearsal for conflict as it will occur in our mature relationships. A natural question one might have then is whether punishment on this view ever constitutes an appropriate response to wrongdoing in the context of mature relationships. Hard treatment is not *all* we rehearse in being punished, but is it one thing we rehearse? Nothing in the Participant View of childhood punishment commits one to thinking that it is, but neither is it ruled out.[54] The Participant View begins from the premise that moral failure is a natural and inevitable feature of our mature, interpersonal relationships, and that wrong-reactive attitudes like resentment will sometimes be appropriately harbored and sometimes appropriately expressed in response. This leaves open the question of whether or not any of these expressions will count as punishments, and what if any value such punishment might serve in the very different context of mature, personal relationships of equality like friendship.

Over the course of this chapter I am going to argue that we (occasionally) do, and (occasionally) *should* punish our friends. The very idea, though, of punishment in the context of

[54] The mature approaches to conflict that we rehearse with children *by* (appropriate/permissible) punishment will not necessarily *include* punishment. There are two reasons for this: First, as I argued in chapter 2, the forms childhood punishment takes are typically exaggerated or modified versions of the sorts of treatment we are training them to practice and to tolerate in adulthood. So, e.g., in putting a child in time-out, we do not teach her that it is acceptable for her to put others in time out when she feels they have behaved badly, or to expect that others will treat her this way in the context of her adult relationships. She may, though, learn to take space in the wake of conflict to calm down. The time-out is a punishment—a deprivation imposed *as such*—intentionally—in response to wrongdoing. The stepping away that (among other things) we thereby learn both to do and to tolerate is not. Or, at any rate, not *necessarily*, as I shall argue here.

a friendship makes many of us uncomfortable, as well it should. In his classic paper "Guilt and Suffering" Herbert Morris gets to what I think is the heart of this worry. "Punishment," he says,

> "is a common response to wrongdoing in non-reciprocal, parent-child relationships and in impersonal, reciprocal legal situations. The role of punishment is non-existent, insignificant, or positively perverse in contexts where moral wrong is done to a stranger or where a friendship or love relationship based on affection, respect, and trust has been damaged…Infliction of punishment by an injured party [in the context of such a relationship] has a peculiar inappropriateness."[55]

Punishment, on this picture, is something rare, strange, and perverse in the context of a mature personal life.

Punishment's perversity here seems to have two faces, the first being that to punish a friend can seem to involve taking oneself to be in some sense her superior. The authority to punish as we typically understand it seems to require the special authority of a judge, a parent, or even a god—all of which stand in some sense *above*. To punish a friend, says Morris, "would sound an unappealing note of moral arrogance," seeming to assume the mantle of special, hierarchical authority inside a relationship of equality.[56] It does not seem to be the place of a friend to judge what suffering one deserve and to impose it unilaterally in the form of punishment. It is likewise not the place of a friend to manipulate one's behavior, even as a means of improving it. Neither does it seem to be the place of a friend to "teach one a lesson" by punishment, as a parent might do with a child. If we understand punishment as aiming at retribution, deterrence, or moral education, it can be hard to see how a peer, with no special form of hierarchical authority, could impose it without fault. Those who do feel entitled to punish on such grounds seem, then, thereby to reveal a failure to understand the basic norms of equality and mutual respect that govern the personal relationships between mature moral agents.

---

[55] Morris 1971, p430.

[56] *Ibid.*

To punish a friend may also seem to involve a kind of practical misunderstanding that is itself perverse. When there is serious betrayal inside a friendship, the appropriate reactions on the part of the betraying friend, and the ones we should want from her, are things like sincere, unprompted remorse, a willingness to reflect on the wrong and to take responsibility for it, and a sense of commitment or recommitment to the relationship and its repair. It is these reactions that have the power to restore the bonds of affection, respect, and trust that betrayal damages. The friend who punishes, then, seems to be making one of two kinds of mistake: either she seeks not these ends, but rather acts on the kinds of vengeful or petty motives that we should take to be inconsistent with an ideal of friendship, or she actually believe that one could restore those bonds of affection, loyalty, and trust, by means of punishment. This would seem to betray a deep misunderstanding of human feeling and value—something akin to thinking that you could make someone love you by locking them in your basement, or that love secured by such means would be worth wanting if you could. One cannot either initiate or restore a friendship by force or violence, and so, whatever valuable work punishment may do in other contexts, it seems to be of no use here.

It looks then like the person who feels justified in punishing a friend is making at least one of three mistakes: She is either (A) taking her normative position to be somehow elevated above that of her friend; (B) taking the norms of friendship to support manipulative, vengeful, or patronizing treatment; or (C) thinking that friendship is the kind of the thing she could force or even *hurt* someone into offering her. Each of these constitutes a very serious kind of mistake. Indeed, one may well wonder whether a person who holds any one of these beliefs could really understand what friendship is at all, or have the capacity to be a good friend. At best, then, punishment in friendship will be the kind of thing one only does against one's better judgment or

higher ideals. In such cases it will not reflect any profound misunderstanding of human relationships or the moral norms governing them, but will constitute, still, some (perhaps understandable) failure to live up to the ideals of friendship under trying conditions. Nothing here, though, worth defending.

There is, I argue, a serious problem with this picture. That problem is not to do with the implicit characterization, here, of what it is to be a good and morally sensitive friend. No reasonable ideal of friendship will include treatment that is petty, vindictive, patronizing, manipulative, or passive aggressive. While we do as a matter of fact sometimes punish in an attempt to change one another, or as a form of mindless tit-for-tat, I don't defend it. Neither do I have any quibble with the notion that any reasonable ideal of friendship will demand that the friend who seriously violates the norms of trust and respect that govern friendship should come to experience some appropriate measure of guilt and contrition. The problem, rather, is with the implicit characterization of punishment itself—of what, fundamentally, punishment is, and what value it serves. If we start from the standard accounts of punishment and its justification, as formulated in the literature on state punishment, then punishment in the context of a friendship would indeed look categorically objectionable. We are not entitled to kill or incarcerate our friends. We cannot send them to their rooms, or suspend their driving privileges. It is inappropriate to punish, in friendship, from motives of retributive justice, or of deterrence, special or general.

Here, though, as in the case of childhood punishment, we ought not import from or treat as exhaustive the range of theories aimed at justifying the state practice of punishment. I have claimed we ought to adopt a general account that treats punishment as a broader phenomenon, which comes in a variety of forms and is common to a variety of relationships. My aim is to

show the possible place and value of punishment in a particular (and ultimately across a variety) of contexts. In chapter 2 I argued that a modified form of moral education theory provides the appropriate grounds for justifying the punishment of children, derived from the parent's more general duty to see to the moral education of children, grounded, in turn, in the even more general duty to see to their wellbeing and development. In chapter 3 I will argue that punishment in friendship is, likewise, justified just to the extent that it serves the constitutive aims of friendship, which is the very different aim of mutual understanding and appreciation. My burden will be to show that punishment is something that is not uncommon in friendship, that friends have an occasional but important authority to punish, and that punishment can effectively work to serve a shared aim of mutual understanding and appreciation, helping to restore the *dynamic* of mutual understanding and appreciation when it has been upset by wrongdoing. Insofar as punishment aims at this restoration in the right way, it *may* serve friendship's purpose, and is on occasion a friend's place.

As in the last chapter, my presentation of the view will proceed in five parts: *Boundary*, *Aim*, *Mechanisms*, *Authority*, and *Permission*. The basic form of the view is this:

a. Boundary. Any intentional imposition by one friend upon another of hard treatment in response to some alleged or perceived wrong will count as an instance of the punishment in friendship. This will include cases in which a friend responds to wrongdoing by temporarily withdrawing from certain aspects of the relationship, where this withdrawal is intended to serve as a deprivation.

b. Aim. The justificatory aim of punishment in friendship is to restore the dynamic of mutual understanding that wrongdoing disrupts. Insofar as friendship is or involves a dynamic of mutual understanding, successful punishment will be restorative of friendship.

c. Methods. Where punishment achieves this aim, it does so by operating as a means of direct communication. It is a way for one friend to communicate to the other, by suspending normal relations in such a way as to deprive the other of the normal experience or expression of the relationship. In this way, she communicates that the relationship has sustained serious damage, and, perhaps, something about the nature and

stakes of that damage. This is a form of communication made possible by shared communicative capacities internal to friendship—made possible, that is, by the fact that friends understand one another.

d. *Authority*. The authority or standing that a friend has to punish is not in virtue (or at least not exclusively in virtue) of being the wronged party, or in virtue of being morally superior, but in virtue of being a *friend*, with responsibilities to the friendship. The friend who punishes appropriately will act as a representative of the friendship, in response to the violation of some norm *governing* the friendship—for, in other words, being a bad friend.

e. *Permission*. Given the nature of punishment's justificatory aim and the particular form of friendly authority, many possible forms of and motives for punishing a friend are ruled out. The form I will defend as occasionally permissible is deprivations of time and attention. These, I will argue, are sometime indispensible tools for the friend invested in real, meaningful repair of a relationship that has sustained serious damage.

Punishment understood in these terms will not in all cases involve the kinds of moral mistakes named in (A)-(C)—the friend who punishes in this way does *not* take her normative position to be elevated, does not treat her friend in a way that is manipulative, vengeful, or patronizing, and does not thereby treat friendship, or contrition within friendship, to be the kind of thing one could force. Once we have the correct view of what punishment is, and what value it aims to serve, the seeming moral problems of punishment in friendship as such will turn out merely to be moral problems with *most* cases of punishment in friendship, but with some important (and illuminating) exceptions.

A. Boundary

In chapter 1, I offered an account of punishment according to which punishment is any imposition of hard treatment in response to wrongdoing. So, the brief answer to the question "what counts as punishment between friends?" is: any instance of hard treatment imposed by one friend upon another in response to wrongdoing. As emerged in both chapters 1 and 2, though, there is room for error, disagreement, and misunderstanding in identifying the kinds of cases to

which that definition might apply. As a matter of conceptual possibility one could attempt to punish a friend by locking her in a basement, and probably someone has. One might also punish a friend by physically assaulting him. This has no doubt happened on many occasions. If, as I argued in chapter 1, any desire, aversion, right, or reasonable expectation can be leveraged to punish, no doubt almost ever desire, aversions, right, or reasonable expectation has been at one time or another used in exactly that way. But there are more familiar and permissible forms of suspending and resuming normal relations in the wake of wrongdoing in friendship. If permissible punishment in friendship does not take the same form as punishment instituted by the state or the parent, what form *does* it typically take?

Consider the case of Aayan and Beatrice—Aay and Bea for short. Aay and Bea are old friends who have seen each other through a lot. In the course of a mundane disagreement, fueled by some combination of underlying tensions and momentary bad temper, Bea says something startling cruel to Aay. She levels an unfair accusation, one that knowingly exploits some secret shame or tenderness. Aay is brought up short. He goes silent. He turns away. For a while he doesn't want to see Bea. He doesn’t want to talk. After these first feelings fade, he resists her attempts at contact and reconciliation for a while longer, despite genuinely missing the solace of her friendship.

Both the cruel way that Bea speaks to Aay here and the silence that Aay treats Bea to in response are typical of punishment as it occurs in our personal, as opposed to institutionally mediated relationships. One says something intentionally hurtful to a friend toward whom one has built up some well of resentment. One withholds the normal expression of affection or support from a friend by whom one feels betrayed. These ways of reacting to perceived wrongs do not always amount to punishments. A cruel remark may be a mere lashing out, like the

flailing of someone startled awake, and not reflect any intent to harm. Silence, likewise, may reflect nothing more than a commitment not to speak rashly when upset. My claim is not that all such forms of treatment will constitute punishments, but that punishment when it occurs in this domain tends to take this kind of form: some temporary loss of the benefits associated with friendship in particular—the benefit of company, attention, conversation, support, or comfort.

To get clear about when these often relatively subtle deprivations count as punishments, it will help to review the criteria for punishment as laid out in chapter 1:

First, *the Wrong-Responsiveness Condition*. For some way that *A* treats *B* to amount to a punishment, *A*'s treatment of *B* must constitute a response to some (perceived or alleged) wrong.

Second, the *Hard-Treatment Condition*. For *A*'s treatment of *B* to count as *hard* treatment, it must, at minimum, seem to constitute some form of deprivation or burden for *B*. If *A* has no reason to hope or suspect that her treatment of *B* will be in some way burdensome or deprive her of something valuable, then whatever she is doing, it is not punishing. So long as some form of treatment *does* seem to *A* to constitute some deprivation or burden, though, it meets the Hard Treatment Condition, even if it is not a material deprivation, or one involving any physical form of suffering.[57] Punishment may, e.g., take the form of withheld attention or a cruel remark.

We are now left with just those responses to wrongdoing that constitute hard treatment. There is still a lot left in this category that does not count as punishing. Say, for instance, that in the wake of some betrayal I decide that I need to have a serious talk with a friend, explaining to him how hurtful it was, and say I know that my friend is extremely sensitive, and will find it very painful to have that kind of conversation. I would be forcing that conversation in response to

[57] I defend this claim at greater length in chapter 1, p22-24.

wrongdoing, fully understanding that this conversation will be painful for my friend. I do not thereby punish her. This is because there is a further condition on punishment—a further, *distinctive feature* of that form of treatment. In punishment, the deprivation or burden cannot be *incidental*. The fact that *A*'s treatment of *B* will burden or deprive *B* has to be among *A*'s reasons for treating *B* in that way. It is the deprivation or burden *as such* that is imposed. This I called the *Imposition Condition*.

When we respond to some perceived or alleged wrong by imposing a deprivation or burden *as such*, we punish. If punishment is held to include more characteristically interpersonal forms of deprivations, then punishment is not (contra-Morris) so uncommon after all. It just looks, as a general rule, much different than punishment as we encounter it in the state case. This is as one should expect given the very different nature of the relationship between punisher and punished. The value of friendship and the value of citizenship are different, as are the duties each involves, and so each renders us differently *responsible* and differently *vulnerable*. How punishment meets the three conditions will therefore vary widely from one kind of relationship to another.

For one thing, punishment will typically (and only permissibly) respond to wrongs particular to the relationship in which it occurs. Some acts will count as wrong in one context but not in the other. I am, for instance, both a friend and a citizen. Driving wildly over the speed limit for kicks might make me a bad citizen without necessarily making me a bad friend. Likewise, breaking a promise to a friend may make me a bad friend without making me a bad citizen.[58] Just as the state is only entitled to punish me (if it ever is) for violations of its laws, so a

[58] It is true that some acts will count as wrong in both contexts—if, say, I speed for fun in violation of some pledge I have made a friend, or at great risk to myself when I know a friend's

friend (as such) is only entitled to punish me (if she ever is) for violations of the norms governing our friendship. Returning to our example case, Aay and Bea each react to alleged failures on the part of the other *qua friend.* Bea, in particular, violates Aay's trust, exploiting her intimate understanding of his vulnerabilities to say something really hurtful. She thereby violates no rule of law, but the norms of trust and care particular to friendship.

It is not only the kinds of wrongs to which punishment appropriately responds that will be different, but the nature of the hard treatment typically imposed. The range of deprivations and burdens even conceptually possible in the context of a loving friendship are much different then those conceivably imposed by the state. The state does not feel or express affection or care in the way an intimate does, and so it is not available to the state to withhold affection or its markers. The state cannot impose an icy silence, or stop bringing you coffee in the morning—not in the way an intimate can. These relationships are *personal,* and so it is possible in the context of a friendship to impose distinctly personal forms of deprivation, each with distinctive forms of expressive power. Though punishment will in any case meet the Hard Treatment Condition, we should not expect hard treatment to take the same form from one context to another.

What it means for punishment to meet the Imposition Condition will also be different when it is imposed by a friend than when it is imposed by the state—or, we might prefer to say, by a person acting in her capacity as a state officer or representative. While it may be more complicated or controversial to attribute intentional action to institutions, it is in practice much easier to identify with at least some degree of certainty whether or not a deprivation is "intentional" in the relevant sense insofar as we can know the institution's aims, which have

---

well-being is crucially dependent upon my own. Even in these cases, though, we have two distinct kinds of wrongdoing in a single act.

been settled by design and express themselves in its daily operations.[59] Consider again this case from chapter 1: Though the DCFS officer who removes a child from someone's home may deprive that person more severely than the judge who sentences him for a drug charge, and though both may be responses to the same exact behavior, we know which is a punishment and which is not. The deprivation imposed by the DCFS officer is incidental—her aim is to keep the child in question safe, not to deprive her parents. The aim of the judge in sentencing those convicted of crimes is, in contrast, precisely to impose deprivations and burdens (though we may disagree about the reasons the state has *for* the imposing).

It will often be more difficult in interpersonal contexts to know with certainty when one is punishing or being punished.[60] This is because in the case of personal punishment, it is not always entirely obvious either to the actor or those observing her what her reasons for acting are. The fact of the matter here is *psychological*, not structural, and introspection has its limits. I may

---

[59] The idea that an institution can be an agent, which acts in ways that might meet the imposition condition is itself a somewhat controversial one, which I will discuss in more detail in chapter 4.

[60] This problem has its complement in a second problem that I will discuss at greater length in chapter 5: Namely, that there is something strange in attributing attitudes or motives to institutions. For it to be true that the state punishes at all, on my definition, it will have to be true that the state does not just impose a deprivation, but that it imposed it *intentionally* (as opposed to its being merely accidental or incidental to the state's aims or workings). There is something potentially odd, though, in characterizing institutions as agents, with "motives". Sharon Dolovich in her treatment of this issue notes that "[i]nstitutions, as complex organizations, lack the unified psychology of natural persons," but does not on this basis think we should conclude that states are not the kinds of things that act or harbor intentions." States, after all, "wage wars, sign treaties, raise taxes, and open or shut their borders to immigration," and they do so for particular purposes (Dolovich 2009, p925). But since the state lacks a unified psychology, we must instead "judge the character of an institutional action, one must therefore look to the institution's design and to the consequences of its operation for those subject to it" (*Ibid*.). Here the idea is that we have a kind of problem when it comes to characterizing the actions of institutions because *institutional* intentional states are not psychological. But, then, the fact that the intentional states of persons *are* psychological do not necessarily make them easier to discern, as the aims of person in any given moment are not formalized in written documents, nor the kind of thing one can always reliably read off of past treatment.

be unsure, in a given case, whether my reasons for storming out of the room include that it would deprive you of something, or how that particular reason weighs against other reasons I also have for acting in that way.[61] We lack, in the personal case, a set of formal indicators that punishing is what we are up to. As with any form of action, then, we have a problem in practice about knowing what our reasons are. We do *not*, though, have a *theoretical* problem about what reasons for action make for an instance of punishing. As in the state case, hard treatment in response to wrongdoing between friends constitutes a punishment just insofar as the hard treatment is purposeful and not mere accident or incident—insofar, that is, as hard treatment is imposed as such.

Let's return now to the case of Aayan and Beatrice. I will leave aside now the question of whether or not Bea's initial treatment of Aay constitutes a punishment. If it does, it is not punishment of the kind I aim to defend—I will stipulate, in fact, that Bea's treatment of Aay was wrong, constituting a kind of betrayal.[62] (If the case as I have specified is not one you can be convinced constitutes a form of betrayal, you may substitute your own.) The question I want to settle is whether or not Aay's response to Bea—his going silent—might constitute a punishment. Here we have a case that by stipulation meets the Wrong-Responsiveness Condition. The

[61] I do not mean to deny that introspection and observation will be sufficient in many cases to determine with a high degree of certainty one's reasons for action, and, thereby, whether or not a particular response to wrongdoing constitutes a punishment. There will be cases where we fail to know our own minds or misinterpret the motives of others, particularly in this arena, where psychic tension runs high.

[62] One may be inclined to deny that speaking cruelly to a friend constitutes a *betrayal*, in the way that, say, a serious, high-stakes lie might. But for an intimate to exploit vulnerabilities of which she is only aware because she has been specially entrusted with the knowledge of them is a rather serious and central form of betrayal in friendship. This strikes me as being true even in cases where the vulnerability in question is not one that one friend has told the other in confidence. Even the act of letting one's guard down in a way that allows another to observe one's vulnerabilities can involve a form of trust that is violated if he should use what he has learned to hurt or manipulate.

treatment that we are concerned to assess is a response to an intimate cruelty—a failure of Bea qua friend. It also meets the Hard-Treatment condition. Bea is deprived, for a time, of certain aspects of the normal experience and expression of that friendship, and the value they have for her. This is also a case that may or may not meet the Imposition Condition. It is perfectly plausible that among Aay's reasons for withdrawing from Bea for a period—and, in particular, for holding this posture even after his initial hard feelings have passed—is that he means to deprive her of some aspect of their friendship for awhile. Insofar as he does, he punishes.

The question that remains is whether there is any way of specifying the case such that Aay *does* meet the Imposition Condition, without thereby necessarily making it the case that Aay is being, however understandably, a bad friend in some respect. Insofar as Aay is intentionally depriving Bea of something, his motives for doing so might be inconsistent with any reasonable ideal of friendship. They might be vengeful or patronizing or manipulative. It might be the case that Aay is over-reacting, treating the breach as being more serious than it really is, or addressing that breach in the wrong way—icing Bea out when he should meet with her and talk it through. These are all possible ways in which Aay, if he is punishing Bea in some way, might well be failing, however understandably, to live up to an ideal of friendship.

The question is whether there are other kinds of cases, too. Is it possible that Aay *meant* to deprive Bea (that the relevant deprivation is neither accidental nor incidental) but did not, thereby, do anything "perverse"? Or is it rather the case that these are mutually exclusive—that either the deprivation is in some important sense incidental or, necessarily, it is out of line?

B. Aim

On the view I advance, the justifying good that punishment of friends may sometimes serve is the good of the friendship, helping to restore a dynamic of mutual understanding and appreciation that wrongdoing disrupts. Where punishment in friendship is permissible, it will serve a communicative function that contributes to the overall wellbeing of the friendship, which is itself a source of value in the lives of its members. On this picture, the defensible version of our case will be one in which Aay's refusal to spend time with Bea, even after his initial feelings of hostility have passed, is intended not to wound or manipulate Bea, but to communicate something to her about the nature and magnitude of the injury she has done to the friendship, and what is at stake. Aay, in acting to make these things clear to Bea, would not thereby act solely on his own behalf, narrowly-conceived, but on behalf of the friendship, in service of friendship's central aims, communicating something to Bea by his withdrawal that is of essential import.

Suppose that Aay does refuse to spend time with Bea for awhile as a means of communicating to her that the friendship has sustained non-negligible damage. (I will discuss at length in the next section how such a withdrawal might work to serve this communicative function.) How does this serve the good of the relationship? My contention is that the wrongdoing itself, if it is a serious enough case to potentially merit punishment, will have damaged the relationship, and that in communicating this fact to Bea, Aay goes some way toward repairing that damage.

To proceed, we will first need to establish something more concrete about what this kind of relationship is, and how wrongdoing damages it. In coming to better understand what kind of value friendship has for us, and what sorts of obligations and privileges it involves, we are better able to say what an *ideal* of that relationship might look like. This ideal, in turn, gives us grounds

for saying, on the one hand, what it is for *B* to wrong *A*, and, on the other, what sorts of responses on *A*'s part might be consistent with that ideal. In the absence of some picture of what friendship amounts to and what it demands, we have no firm basis for saying what the value of punishment might be in that context, or what authority one might have in the context of that relationship to punishment. So just as in the last chapter I appealed to a more general picture the aims and obligations of the parent to ground my account of the value of childhood punishment and the special authority of the parent to punish, so, here, I turn to a more general picture of the aims and obligations of friendship.

Friendship as I understand it is a norm-governed dynamic of mutual understanding and appreciation.[63] Put more simply, to be friends is, among other things at least, to know and care about one another, and to *owe* one another certain kinds of care, attention, and disclosure.[64]

---

[63] In the current literature, as I understand it, friendship is variously conceived of as a form of special concern, a form of mutual love or caring, or a form of intimacy, where intimacy, in turn, is understood in a variety of ways. Among those philosophers who have understood friendship to be a form of intimacy, some have understood intimacy to consist in a practice of mutual self-disclosure (Thomas 1987, pp89, 93), while others have understood intimacy to amount to solidarity (White 2001) or like-mindedness (Telfer 1970), others a mutual receptivity to influence (Rorty 1986, Cocking and Kennett 1998), and still others a singleness of mind involving anything from empathetic identification to plural agency, with joint cares, desires, actions and evaluative perspectives (Sherman 1987, Helm 2008).

The view I begin from here is my own, advanced in greater detail elsewhere. It is an intimacy-type account of friendship, but one on which the relevant intimacy is generated by *a shared aim of mutual understanding and appreciation*. On this account friendship will often and sometimes even necessarily involve those forms of concern, caring, and intimacy around which other theories of friendship have centered. But friendship's ultimate aim—its *distinctive* aim—its *constitutive* aim, is a shared one—the shared aim of mutual appreciation and understanding.

[64] This requirement on friendship is not as strong as it may first appear. Not all friendships involve the same degree or dimension of knowing or understanding one another, and no friendship involves knowing or understanding one another entirely. We come to understand and appreciate one another not to any particular overall degree, but, rather, *in certain respects*. We might, for instance, be friends only qua New York Giants fans, and know and appreciate one another only as such. This is a kind of friendship, and one that does in fact involve a special intimacy and understanding. Such friends may well understand one another *qua* Giants fans

When I say that this knowing and caring is *dynamic* and *mutual*, I mean to capture the fact that for us to be friends is not merely for there to be a two-directional flow of understanding and appreciating, in which I understand and appreciate some things about you, and you understand and appreciate some things about me. Each of us might, after all, secretly be observing and admiring the other without it being the case that we are friends. We might be watching through our binoculars from across the street, the one never catching the other. In friendship, by contrast, we *share* our understanding and appreciation with one another. It is not just that each of us knows and appreciates the other. Each of us knows *that* we are known and appreciated, so that among the things I know and appreciate about you is that you know and appreciate some things about me and vice versa. And each of us knows, in turn, that our understanding and appreciation of the other is itself understood and appreciated. This mutual understanding is governed by norms of both disclosure and attention, which regulate the ways in which we communicate ourselves to one another. Friendship seems to demand at least a minimal reciprocity in the ways we disclose information about ourselves, and attend to information about the other. Mutual understanding is governed, too, by norms of, e.g., trust and respect, which are what make it safe and possible for us to open ourselves up to being known, and to come to rely on one another as sources of support.

---

better than others with whom they are, overall, much closer. I may be friends with someone as a colleague, and our knowing and appreciating one another may be contained within certain profession boundaries. These are not lesser friendship or friendships falling shy of an ideal of perfect understanding, just friendships of a particular kind or character—*work* friends, *childhood* friends, fishing buddies. What makes a friendship is not the particular respect in which two people know and appreciate one another, but that this knowing and appreciating is mutual.

Likewise, the norms governing any particular relationship, including obligations of care and attention, may be relatively minimal, or relegated to particular areas of our lives. So, e.g., if you and I are friends *qua* New York Giants fans, probably what obligations of care, attention, and disclosure I have will be limited to those with some bearing on how we know and act around one another in our capacity as sports fans.

In an ideal of friendship we, as friends, may not know one another in all the same respects or to the same degree, and there may be differences in our relative ability to live up to the norms of trust, respect, and so forth, which govern our friendship. Still, insofar as we are friends, we will at least share a sense of how much we do know and care for one another, and of what the norms governing our relationship are. We share, in other words, a sense of what our relationship is, and of what would count as a violation of its terms.[65] [66]

When we violate the terms of a friendship in sufficiently serious ways, we do damage to the relationship itself, independent of the damage we may directly do to the person who is our friend. Bea's words may well have hurt Aay's feelings, but this is not itself a violation of the terms of their relationship. After all not just anything Bea might do to hurt Aay's feelings is something Aay, as Bea's friend, has a right against. Bea may, e.g., owe Aay an honest account of her feelings, though she knows it will hurt him to hear them. It may be precisely *as* Aay's friend that she owes him this accounting. Insofar, though, as Bea acts in such a way as to violate the terms of the relationship, she hurts not just Aay, but the relationship, too. This is what has happened in our case, where Bea's comments not only hurt Aay's feelings, but constituted a violation of the norms of trust that govern their friendship. Their shared sense of trust is part of the framework of the relationship itself, and trust has been broken.

---

[65] I do not claim that those who misunderstand one another—whose views of how well they know one another or what the rules of their relationship are may differ—are not really friends. I only claim that such misunderstandings constitute a kind of impairment of friendship.

[66] This is an ideal of friendship. Studies suggest, in fact, that though we almost always expect that others reciprocate our own sense of how close our relationship to them is, this only turns out to be true about half of the time. In nearly half of our friendships, in other words, we are mistaken in our belief that the feeling is mutual (Almaatouq et al. 2016).

A second sense in which wrongdoing damages friendship is the way in which it can disrupt the dynamic of shared understanding and mutual appreciation that at least partly constitutes that relationship. Sometimes alleged wrongdoing reveals that we did not have the shared understanding we believed ourselves to—as in a case where, e.g., some behavior of yours violates a norm of conduct that I thought we agreed governed our relationship, but that, it turns out, you do not take to govern our relationship. In a case like this, we come to discover that we in fact disagree about whether or not you have done anything wrong. In other cases we will disagree, instead, about whether some behavior of yours or of mine amounts to an actual violation of the norm we both agree governs our friendship, or how serious a violation it constitutes. Very often in the wake of wrongdoing the wronged party will feel that his wrongdoer fails to appreciate the nature or seriousness of the violation.[67] This disagreement about so central a thing as whether or not one party's treatment of the other constitutes a serious violation of the terms of their relationship *itself* constitutes a disruption of the dynamic of mutual understanding and appreciation in which friendship consists.

What punishment in its communicative function can sometimes work to achieve is to get friends back on the same page. It can function as a means of letting someone know that they have done serious damage to the relationship—or, at least, of registering the fact that one party takes this to be true. Where punishment succeeds in making someone understand this fact, or

[67] These disagreements can sometimes work the other way, too, with the wrongdoer feeling that the friend she has wronged has failed to grasp how serious his violation was. These types of cases may be less common, but they do occur. The wrongdoer is, after all, in some ways better positioned to understand with some depth and subtly precisely what he did. This kind of failure of shared understanding presents its own problem for friendship. The wrongdoer may struggle with the question of whether or not it is worthwhile to go to the effort of trying to make his friend understand, and may struggle with feelings of shame that can make it difficult to make such an effort, even if he thinks he should. These are not the kinds of barriers to mutual understanding that punishment works to address.

advances a conversation in which this fact is eventually established, it works directly to restore friendship in this second sense—helping to reestablish the dynamic of mutual understanding and appreciation. It can also play a role, albeit more indirectly, in reestablishing the norms and bonds of trust that genuine wrongdoing breaks with. This form of repair tends to require contrition on the part of the wrongdoer, which begins with a non-voluntary feeling of guilt, and a process of coming to identify with, or develop attitudes of acceptance toward, that guilt, and a willingness to work at making amends. The very initial stages of this process, though—the non-voluntary guilt reaction—has to begin from an understanding of what one has done. It is this very understanding to which the non-voluntary guilt, if it is appropriate, will constitute a reaction. It is your sense of our shared understanding of what happened that lends any subsequent contrition or apology of mine its normative weight.

In sum, then, Aay, if he is acting permissibly, wants Bea to know that her behavior constitutes a serious breach of trust, and that their friendship has thus sustained a kind of damage. In acting to bring it about that Bea knows this, Aay serves the friendship in two ways: First, insofar as Aay succeeds in communicating this to Bea—that is, insofar, as Bea comes to understand and appreciate what he is telling her—he gets them on the same page. This in itself constitutes a partial restoration of the dynamic of mutual understanding and appreciation essential to friendship. This shared understanding of the damage done to the relationship, in turn, makes further reparative steps like contrition on Bea's part possible. In this way, Aay serves the good of the friendship by acting to communicate information, the communication of which is itself partially restorative of the relationship, and which makes further, even full restoration possible. He thus acts, in punishing, for the sake of the friendship.

C. Method

I have said that the suspension of normal relations in friendship can constitute a kind of punishment, so long as it is both a response to wrongdoing, and intended to deprive or burden. I have also said that punishment of this form can work to communicate information in a way that both constitutes a kind of repair, and that makes possible repair of a further kind.[68] What I still owe is an explanation of how punishment, so understood, could work to serve this aim. These are, after all, somewhat counterintuitive claims: That walking away, cancelling plans, *not* talking, *communicates*; that acting to deprive someone of your friendship could serve to improve or repair that friendship. How?

There is nothing new, of course, about expressive theories of punishment. Retributivists and deterrence theorists alike acknowledge that one of the ways in which punishment operates is to express the community's condemnation of wrongdoing, its resentment and indignation toward the wrongdoer, and, on some accounts, its conviction that the victim of wrongdoing deserved better.[69] But accounts of punishment's expressive dimension have often failed to include any clear articulation of how, precisely, punishment works to communicate. Punishment on such theories is said to have some "symbolic significance" or "conventional meaning".[70] Hard

---

[68] I do not here deny that other forms of punishment may be permissible, nor that punishment may sometimes be permissible for other reasons. My aim is simply to show that punishment in *some* form may be permissible, and so I focus on this kind of case in particular, which seems to me most defensible.

[69] See especially Feinberg 1965, Hampton 1984, Hampton 1992, Duff 2001, Lacey 1988, Primoratz 1989, Braithwaite & Pettit 1990, von Hirsch 1993, Boonin 2008, and Kleinfeld 2016.

[70] Feinberg 1965, p402.

treatment is held to *mean something*, having "become the conventional symbols of public approbation," in much the same way "that certain words have become conventional vehicles in our language for the expression of certain attitudes."[71] The analogy to written or spoken language here is strong, and perhaps explains why very little direct attention is given to explicating how punishment works to serve its expressive function. We are to understand that it functions in much the same way as any other form of communication: We have come, as a matter of convention, to express particular attitudes in a particular set of terms, and when we "speak" in these terms, we reach the people we are "speaking" to by engaging them at the level of a set of rational communicative capacities. It is a message that anyone can understand, so long as they are possessed of a capacity to grasp the concept of wrongdoing and sufficient familiarity with the local conventions for expressing our sense of it. Where philosophers concerned with punishment's expressive function have not leaned on the idea of punishment's meaning as being a matter of shared convention, they have tended to appeal to a kind of *natural* meaning, claiming that punishment is a kind of "natural" expression of the wrong-reactive attitudes, which we perhaps understand just in virtue of being the kinds of creatures who also experience those attitudes.[72]

Moral education theories of punishment in particular tend to have a strong expressive component, and have tended to address more directly the question of how punishment communicates. This is part because most contemporary moral education theorists are concerned to distinguish particular means by which punishment "educates" from others. Jean Hampton, for

---

[71] *Ibid.*

[72] A.J. Skillen has called this "'the most natural-thing-in-the-world' justification"—"Like many of its kind," he says, "it works in part by pretending that specific forms of punishment, for example capital punishment, are the natural expressions of condemnation" (Skillen 1980, p514).

instance, distinguishes between two kinds of messages that punishment sends: "If one wants someone to understand that an offense is immoral," she writes,

> "at the very least one has to convey to him or her that it is prohibited—that it ought not to occur. Pain is the way to convey that message. The pain says "Don't!"…an animal shocked by a fence gets the same kind of message…But the state also wants to use the pain of punishment to get the human wrongdoer to reflect on the moral reasons for that barrier's existence, so that he will make the decision to reject the prohibited action for moral reasons, rather than for the self-interested reason of avoiding pain."[73]

The first level at which punishment communicates is to deliver a simple "Don't! (or *else*)". The second, "*moral* message implicit in punishment" is that there are (and the object of punishment *has*) moral reasons not to engage in the kind of behavior in which the wrongdoer has engaged.[74] The first kind of "message" is one, Hampton points out, that almost any kind of animal can receive. The second kind of message, by contrast, is pitched at the level of rational moral agency. The autonomous moral agent, capable of understanding this message, is then free to reject or take it on board as she sees fit. Punishment, in short, says "*That kind of behavior is wrong—you have moral reasons to refrain from it*," and the punisher can then choose to "listen" or not.

This is a significant distinction for Hampton in part because she thinks that without it, we run into a problem about justification. While the state, she thinks, has a right and a responsibility to ensure the safety of those who live under its laws by deterring law-breaking, it is not entitled to treat human beings like animals, "using pain coercively so as to progressively eliminate certain types of behaviors."[75] Because punishment "says 'Don't!'" to the wrongdoer in the same way an electrical fence says "Don't!" to the livestock, Hampton thinks we can only justify its use if it aims instead to communicate to wrongdoers the *moral* message, which does not have its

---

[73] Hampton 1984, p212.

[74] *Ibid*. emphasis added.

[75] *Ibid*. p214.

impact by a process of animal conditioning, but by engaging the reflective capacities of the wrongdoer. If punishment is to be consistent with basic respect for person, then our aim in punishing must be to deliver a moral message, which will move the wrongdoer (if he chooses to listen) by way of his capacity for rational, moral reflection.

I am offering a different kind of expressive theory—one that addresses the question of how punishment works to communicate, while remaining sensitive to Hampton's distinction and its moral import. First, on these standard expressive theories of state punishment (moral education and otherwise), the expressive function of punishment is to communicate to the wrongdoer (and to everyone else) that his act was wrong, disavowing such action, and expressing a form of solidarity with the person wronged. On an analogous view, brought to bear in the context of a friendship, we might understand the meaning of Aay's withdrawal in something like the following terms: "You, Bea, have wronged me. I don't approve of how you've treated me, and I respect myself too much not to say so." I will not comment here on the moral status of punishment that, whether by authorial intent or an accident of convention, expresses these sentiments, except to say that it is not the conception of punishment that I am concerned here to defend. The expressive aim of punishment that I have been concerned to describe is not to tell the wrongdoer that she is bad, and that her behavior is unacceptable, but to signal to her that the relationship has sustained damage and requires attention.

My theory also differs from the standard theories in terms of how it works to communicate its message. On accounts like both Feinberg's and Hampton's, the language in which we communicate the relevant message is one of pain or suffering. It is the imposed pain (physical or otherwise) that says "Don't!" The particulars of the pain and suffering by which punishment "speaks" are then a matter of either conventional meaning—"almost an arbitrary

code"—or by some alleged "natural" meaning, according to which, e.g. incarceration, or community service, or the pillory, either just is or somehow comes to be the vehicle whereby the message is delivered.[76] Pain, here, is the method, and it is meant to work by engaging the capacity for reflective understanding that distinguishes human beings from mere livestock.

On the view I offer here, by contrast, is not necessarily by the imposition of pain (of whatever kind) that the punisher "speaks", but the imposition of a deprivation of something of value—namely, the normal experience and expression of an intimate relationship. This deprivation may reliably *cause* pain—indeed we may have trouble fully distinguishing in practice between the loss of a valuable relationship and the feelings of bereftness that partly constitute our recognition of a loss like that. It is the deprivation itself, though, and not the associated suffering, that constitutes punishment. There will be cases of punishment in which the deprivation imposed is *itself* a form of physical or psychological deprivation, but this is not the sort of case I aim to defend here, and it is not sort of case I have described. The thing of which Bea is temporarily deprived is not her emotional equilibrium, but certain valuable aspects of her friendship with Aay.

One thing this expressive view shares with the others (or with Hampton's at any rate) is that it claims that insofar as punishment achieves its legitimizing aim, it does so by engaging the communicative capacities of moral agents. This kind of claim is rendered much more plausible, though, in the context of friendship—especially intimate friendship. Consider for a moment the general question of how communication typically works in a friendship. One way we do it is by verbal reporting—e.g. "The Giants won last night"; "My first memory is of falling"; "I like you." This way of telling you things may be particularly important if we are new friends. We

[76] Skillen 1980, p511

communicate by other means, too. Take the following kind of case: I communicate to you by an exaggerated eye roll, that the person I am on the phone with is tedious. You are not just reading my tedium here in the way a poker player reads an opponent's nervous tick. I am telling you something. I am communicating something to you—something I know you will understand as surely as if I said to you, "This person is tedious."

As friendship grows, so does our shared capacity for non-verbal communication. You know more about the idiosyncrasies of my gestures and expressions and I know you do, and so we share an even wider vocabulary of gesture and expression. In this way, and many others, we develop over time a shared language between us—a means of communicating that, while not always verbal, is nonetheless direct—that is to say, not just meant to be read in a certain way, but a kind of *telling*. Some of the attending I do as your friend involves noticing things about you that you have not, in this sense, "said". I might notice, e.g., that you are upset, perhaps before you yourself have noticed it. Some of the attending involves listening to you offer verbal reports. But such communication often happens (perhaps even a great majority of the time) by means of non-verbal communication in its myriad forms, reliant on the capacities of friends to express themselves in and to grasp the expressions of the other in this non-verbal language. These communicative capacities develop as the friendship develops. Friendship, involving as it does these forms of mutual understanding, by its nature expands our communicative capacities. They are woven into the complex network of mutual understanding in which friendship consists.

Of particular importance to this picture of punishment and how it communicates is the question of how friends communicate to one another the value that their friendship has to them. The dynamic of friendship as I have described it requires that you know something about my appreciation of you, and that I value our friendship. If I care about our friendship, and I want you

to know it, I may manage this by means of verbal reporting, but that is only one of many ways of making our appreciation and valuing known, and one that we can very often do without. More commonly I express my valuing of the relationship by my investments of time and attention. All investments of time and attention may in this way *demonstrate* that I care. but in friendship I can also *use* certain investments of time and attention *tell* you I care. What makes this possible is our mutual understanding that this is a way we demonstrate care. If *I* know that *you* know that *know* that this is one of the ways that we demonstrate care, and vice versa, then something more than mere expression-as-demonstration is possible and, on occasion, a wonderful thing. Say, for instance, that I am in the final weeks of finishing a dissertation, having neglected you and everything else for months. There's just no time. But say, too, that I come anyway to be with you in a moment that I know is important to you, spending time with you we both know I don't have. Having missed so much else, I want you to know that I value our friendship—I want to show you, sure, but I also want to tell you. With a less intimate friend, I might not be able to communicate in this way, but I know you understand me, and that when I walk in and we see each other, I trust that you'll know what I mean.[77]

[77] There is a way of reading this, and potentially all such instances of "telling" as sinister, or at least distasteful—like the person who is acting kindly for the purposes of making people think that she is kind, rather than as a natural response to the features of her situation that call for kindness—a phenomenon often called "virtue signaling". This is, in one sense, just not the kind of thing I have in mind. This is, rather, just another way of *demonstrating*, where what does wrong is that the actor is aiming to directly to demonstrate kindness, rather than demonstrating it incidentally, on the way to *being* it. Telling is something else. And at least on occasions like the one I have described here, it is not a distasteful instance of virtue signaling. It is, rather, a way of employing the especially intimate communicative capacities of friendship to say something that we, as friends, ought to find ways of saying sometimes. It is true that we ought to *show* that we care, and that mere "telling" in the absence of showing is hollow. This does not, though, mean that *all* telling is hollow, or that we are not also required or, surely, at least, permitted, to do some telling.

This sort of telling can have some distinct advantages over telling discursively. I will discuss this matter further in the section on permissibility, but for now I simply note that this form of communication does, in fact, often have a unique impact. Spoken language has its limitations, and we feel them deeply. Speaking by our actions not only seems to be a way of making our meaning more clearly understood, but making it more clearly understood that we mean it.

If we can directly communicate (telling, rather than merely showing) something about the friendship and our attitudes toward it by means of showing up and spending time together, we sometimes communicate in the same way by *not* spending time together. Sometimes an absence, like a presence, simply *shows*. That we have stopped reaching out to one another might simply show that we have grown apart. But withdrawals can also be a way to *tell*. Aay, I am claiming, may withdraw, as a means of telling Bea something—by means he can trust that Bea (as a friend) will understand. (He trusts in this because he knows they share both a sense of the relationship's value, and an understanding that in, e.g., spending time together and not spending time together, they can communicate about that value.) Aay would be aiming, thereby, to communicate that there has been damage done to the friendship, and perhaps, depending on the depth of their mutual understanding, even something about the extent of the harm, and the stakes involved (the value of the damaged friendship, to both friends). These facts are communicated, as they must be, inside the dialectic of friendship—in the context of this dynamic of mutual understanding and appreciation. It is a context in which non-verbal communication may be as direct as any verbal communication, and in which our shared understanding of the expressed meaning of time spent together transforms the ways in which we do and do not spend time and attention on one another

into a possible means of communicating with one another about the state of the friendship. In the context of a friendship—of mutual understanding—we have special communicative capacities

A virtue of this understanding of the mechanisms by which punishment works to serve its justifying aim is that it avoids the legitimate worries of moral education theorists like Hampton that punishment might work by treating people like animals to be conditioned, or instrumentalizing their rational capacities as a means of manipulating their behavior. In "saying" by withdrawal, we neither bypass nor manipulate the rational capacities of the other, but engage them directly. And in this case, unlike the case of state punishment, the intimate and communicative nature of the relationship renders the notion of a shared non-verbal communicative capacity between punisher and punished more full and plausible. If Aay's treatment of Bea is effective in its permissible aim, it will not merely trigger some set of beliefs or behaviors in Bea that Aay would like to see. He is not dropping hints he hopes she will pick up on. If Aay is successful, Bea will come to understand something she could not or did not before about what she has done, and she will have come to understand it because Aay, in walking away, stood up for the friendship and told her.

D. Authority

Even for those inclined to be sympathetic to the picture I have sketched of punishment's occasional value in friendship, there might well seem to be a real problem about the *authority* of a friend to punish. Putting punishment aside for a moment, the very notion of authority can seem to involve some idea of a right to control, demand, and even force obedience. Authority is often said to be something we have "over" one another. These notions of hierarchy and control are antithetical to any reasonable ideal of friendship, which, though it supports much difference, is a

relationship of equality. It is also a *chosen* relationship. Most of the relationships in which we acknowledge an authority to punish, in particular, are more or less insoluble. Neither children nor, in most cases, citizens have either the right or the wherewithal to dissolve their relationship with the punishing authority. Friendship, in contrast, is meant to be the kind of relationship into which we enter freely, and that we are, with some complications, generally entitled to leave. Friendship, unlike the relationship between citizen and state or parent and child, seems never to tolerate coercion, and to be fundamentally symmetrical in terms of the basic orientation of the parties toward one another. What, then, might authority to punish look like in this context?

A good place to start in answering some of these worries is to deflate the notion of authority a little. To be someone's friend is to be specially entitled to them in particular ways, and it is also to be specially entitled to *treat* them in particular ways. Say, for example, that you are going through a very difficult time, and so for awhile you tromp around the department most days being more or less awful to everyone. And say, too, that all but one member of the department are merely acquaintances or rather casual friends, while the one remaining member is a close friend. And say, finally, that everyone in the department is aware that (1) you are going through a hard time, (2) you are being awful to everyone, and that (3) you yourself would probably feel a lot better if you made a concerted effort to engage in more pleasant interactions throughout the day. If I were one of your acquaintances in this case, I might say to myself something like "it would be better for them if someone were to make them understand that they are being awful, and that they would probably feel a lot better if they made a concerted effort to have more pleasant interactions throughout the day, but it is not *my* place to try to make them understand it. It would be out of line to say something" And if, by contrast, I were your one close friend, watching you tromp around being awful to everyone, I might say to myself instead, "as

their friend, it my job to say something here." It is sometimes the place of our friends to tell us hard truths about ourselves—truths that it might be totally out of line for others to confront us with.

When I ask what "special authority" a friend has to punish, what I mean to ask is: in virtue of what is it a *friend*'s place to impose the sorts of deprivations in response to wrongdoing that, according to my account, constitute punishment? One initially plausible answer to this question might be that it is the friend's place to punish in the kinds of cases under discussion because it is the friend who is the injured party. But on the picture of punishment I have offered, the friend who punishes permissibly is responding not to some injury done directly to *himself*, but to an injury done to the *friendship*. He acts, then, not as a representative of his own personal interests, but as a representative of the relationship, and his "special authority" to act as representative, or "on behalf of" the relationship is as a member or citizen of that that relationship, invested with some responsibility for its wellbeing. It is not, then, *qua* injured party that the friend has the authority to punish, but *qua* friend.

The form of this answer to the authority question, then, is analogous to the answer given to the authority question in the account of childhood punishment: In that case the parent has special authority to punish because it is the parent who is responsible for seeing to the well-being of the child, which is the aim that punishing in that context serves. A friend, likewise, is entitled to punish because it is the friend who is responsible for seeing to the well-being of the friendship, which is the end that punishment serves in *this* context. But here is where the two contexts differ in terms of authority relations: In the case of childhood punishment, the aim that punishment serves is one for which one member of the relationship (the parent) is entirely responsible. In the case of friendship, by contrast, the responsibility is shared. Both friends, equally, are citizens of

the friendship, and they share the responsibility for its wellbeing. Why, then, in this case, is it one friend and not the other who is authorized to act as representative of the friendship? The natural move here would be to say that the relevant distinction between the normative position of these two friends is that one (the one authorized to punish) is the injured party. I have denied, though, that being the injured party is what confers the relevant form of authority here. What else could it be?

The answer here, I think, is that there is, in fact, no difference in the form of authority that each friend has. Both Aay and Bea have, as friends, the responsibility of acting in service of the friendship, where in this case that means, among other things, coming to share an understanding of the wrong that occurred. But where one party is entitled to punish, it will be for two straightforward reasons: First, because as a conceptual matter, nothing that *B* does to *A* will count as punishing, because it will not constitutes a response to some wrong *A* committed. *A* has committed no wrong—or at any rate the wrong committed is not the one to which her act will constitute a response. It is responses to *B*'s wrong that we are considering here. It may be available to *B*, in the wake of her wrong doing to *self*-punish, but not to punish *A*. It is, of course, a conceptual possibility that *B* will act in the wake of her own wrongdoing to *deprive A* as a way of communicating to *A* something about the wrongdoing to the friendship that *B's* own wrongdoing caused, even if we don't call that punishment. Perhaps this *will* turn out to be something *B* is allowed to do on those grounds, although it is harder to see how (at least in most standard cases) *B's* depriving *A* could serve the relevant purpose here.

To say that one friend and not the other is in the appropriate normative position to punish, or seems to be, is not to claim that this same person is the ultimate authority *on* his own authority. It may be the case, for example, that Aay believes himself to be entitled to punish but

is wrong—perhaps because he is mistaken in his belief that Bea's behavior constitutes a breach of trust and not just a case of hurt feelings. Or it may be the case that Aay is right that there has been a breach of trust, but that his punishment of Aay stretches past what is needed to communicate the relevant information. It may well be that in such cases Bea is better positioned than Aay to see that this is so, and is therefore obliged, or at least entitled, *as* Aay's friend, to push back against Aay's sense of what he is, as a friend, he is entitled to do here. There is, in other words, a place for both parties in a proper negotiation of what responses to wrongdoing are appropriate.

While Aay's authority to punish is not exclusive, in that it is an authority that is in an important sense *shared* (with Bea), it *is* exclusive in the sense that it is not shared with others. So, for instance, Aay may have other friends who observe Bea's wrongdoing, and who may, being good and sympathetic friends to Aay, even feel this wrongdoing deeply. But this is does not entitle Aay's other friends to intervene and punish Bea on his behalf. Nor is Aay entitled to punish Bea for, e.g., being a bad mother, or a bad student, or for being ungrateful to her parents. The authority to punish here is grounded in the terms of the very relationship whose terms have been violated.[78]

[78] This is in contrast to the case of parental punishment, in which the parent punishes the child not only in cases where the child does some harm to the parent, but also where the child does harm to others, or even violates certain prudential rules (e.g. "no cookies before dinner"). This is a special feature of parental punishment, which is importantly distinct from the sorts of redress that will occur in other relationships. The *justificatory aim* of punishment in the parent/child case is not to repair of the parent/child relationship (which is by its nature the kind of relationship that tolerates typically tolerates such harms), but rather to teach the child how to engage in repair in the context of a whole array of relationships, both personal and institutional, in which she will, if all goes well, eventually participate as a mature agent. It is still the case, though, that the parent's authority to punish is grounded in the terms of the relationship he stands in *to the offender* (in this case, his child). In this case, though, the terms of that relationship allow him to interfere in a wider variety of domains. It may well be the case that the parent of a very young child is entitled

E. Permissibility

Granting that it is possible for Aay to engage in this form of communication-by-withdrawal one might still hold that Aay is not permitted to engage in it. Sure, it is available to Aay to communicate in this way with Bea. What, though, entitles him to pursue that valuable communicative aim in this particular way? Why resort to punishing when one might communicate instead by way of a conversation?

The first thing to note here is that there are many forms of punishment that we are not on this account entitled to impose on one another in the context of our personal relationships, however effective they may be as a means of communicating an important idea. We are not, for instance, entitled to violate any basic right a friend may have *not* in virtue of status as friend, but in virtue of being, e.g., a fellow human being (thus the now well-established prohibition against locking each other in basements). The only form of deprivation I have defended as a potentially permissible means of punishing a friend involves the withholding of that which is only owed (and "owed" here is perhaps even too strong a word) to someone in virtue of their *being* a friend. I took as my example a case in which resources of time and attention are withheld. There may be other defensible kinds of depriving or burdening imposed in the context of a friendship, but I suspect that none of them will involve being deprived of something to which someone has a right, or burdened in some way one has a right *against*, where this right is based in some status of theirs other than the status of friend. I cannot permissibly punish a friend some way that would violate their rights as a person, parent, citizen, or spouse. We are only entitled, in punishing

---

to punish that child for hurting a friend, though the (mature) friend is *not* entitled to punish her friend for being an ungrateful daughter.

friends, to withhold (at least some of) those things which one only owed in virtue of *being* a friend to begin with.

What's more, on this view even our entitlement to withhold *friendship* as a means of punishing a friend is limited. It is first of all limited in the sense that we are not permitted to do so in ways that will fail to serve the legitimate communicative aim of such acts. But there is a further, in-principle limitation, too: One is never entitled, on this view, to *end* a friendship as a means of punishing a friend. It is only the partial and/or temporary withdrawal from the normal rhythms of friendship that I have defended. It is a central feature of this view that the aim of punishment is reparative, and the authority to punish is something one only has in virtue of being a committed friend, with a responsibility to the relationship and its maintenance. This authority is not one that could license punishment in the form of terminating the very relationship in which one's authority to punish is grounded.[79]

Even after having carefully narrowed the question of authority to focus on the particular forms of punishment I have attempted to defend, we can still ask, though, why attempt to serve this particular communicative, relationship-centered aim in *this* way? Why *hard treatment*? Why

[79] I do not claim that it is never permissible to punish interpersonally someone who is one's friend by ending the friendship. I claim, rather, one is never permitted to punish interpersonally someone who is one's friend *as* a friend. The importance of this distinction comes out when we notice that we generally stand in more than one kind of relationship with the people we are know—certainly with those with whom we are intimate. Everyone who is my friend is also a fellow human being. Some of the people who are my friends are also my colleagues. Some of the people who are my friends are also my sister. It may be the case, then, that I sometimes have to engage with the person who is my friend not as a friend, but as a colleague, a sister, or a fellow human being. So perhaps it is the case that a person who is my friend could wrong me not as a friend, in violation of the norms of friendship in particular, but as a person, in violation of the basic norms that govern this more general relationship. And it might, then, be the case that I have standing, *qua* fellow human being, attempting to communicate something to you about the state of our relationship as fellow human being, to end our *friendship* as a means of doing so. I am not sure about this case, but it at least not ruled out in the way that ending a friendship *qua friend* most certainly is.

secure your valuable aim by means of *punishment* of *any* kind if there are other means available to you? If the aim is communicative, why not just have a conversation? One ought not, of course, underestimate the potential pain and discomfort of such conversations, or fail to note that there is nothing on this account that excludes the possibility that conversation will be important or even essential to and sufficient for a successful reparative process in many or even most cases. Still, this is an important question, and one that expressive theories of punishment have not always been able to answer in a satisfying way.

My first response here is to say that punishment, as I have described it, *is* part of a conversation, and not an alternative to it. It is, as I have described it, a way of "telling" and not necessarily any less direct a way of telling a friend something than a verbal report would be. We might even be inclined to strengthen this claim to reflect a sense that actions may even, in this case "speak louder than words." There is a second kind of response one might offer, though, which is that we may sometimes, in punishing, say something that could not have been properly communicated in another way. I have said that in punishing permissibly, we act as representatives of the friendship, "speaking" on its behalf. I have also argued that things like attention, disclosure, and shared activity are means by which we express and experience the value of friendship in our lives. Another way of putting this is to say that friendship's "language"—its vocabulary—just *is* investments of time, attention, trust.[80] That makes this form of communication-by-action not just a matter of conventional meaning established by long practice of interpersonal communication between two particular people, but something more natural and necessary. Punishment may, in this sense, turn out to be indispensible—not just *one* of the ways we *happen* to deliver the essential message, but in sufficiently serious cases the

[80] I am grateful to Barbara Herman for helping me to see this point more clearly.

singular means available to us. If this is so then here, as in the case of childhood punishment, it may well be true that while our sense of what *sorts* of punishment count as permissible may change, we will always need something that meets punishment's three conditions as part of our shared grammar of wrongdoing and redress.

Conclusion.

I began this paper by trying to articulate the discomfort many of us have with the idea of punishment in friendship, settling on three central worries: That punishment, where it amounts to anything more than bare lashing out, would necessarily turn out to be a case in which the person doing the punishing either (A) takes herself to be her friend's superior rather than her friend's equal; (B) takes the norms of friendship to support manipulative, vengeful, or patronizing treatment; or (C) thinks that friendship is the kind of the thing she could force or even *hurt* someone into offering her. The case of punishment in friendship I have defended is a case in which the punisher in question makes none of these fundamental moral errors. He acts *as a friend*, from a normative position that while *authoritative*, is importantly symmetrical and non-hierarchical. He acts *not* from vengeance, nor to manipulate for self-interested ends, nor, paternalistically, for the sake of his friend's personal wellbeing. There is, I have argued, another kind of aim and motive available: Aay can punish for the sake of the friendship. And, finally, Aay has not, in punishing, acted in a way that is either itself coercive, nor aimed at forcing a commitment to the friendship. Aay does not "force" Bea to do anything, except, perhaps, in the sense than I "force" you to hear something you would prefer not to hear merely by choosing to say it to you.

Chapter 4: Grace My Fear Relieved: An Alternative (to a) Theory of Self-Punishment

A. Introduction

Over the course of the last several chapters I have considered the potential value of punishment between persons, and its place in the process of moving on together inside of some particular kind of relationship. [81] I have argued that punishment can play a (limited and occasional) role in helping us appropriately to express and learn to express a set of wrong-reactive attitudes, to come to terms with the wrongs we have done or been done, and to repair the relationship that carries us forward. There is another question in the neighborhood though that's worth considering: What about *intra*personal punishment?

This may seem like a form of punishment that puts the lie to my claim that a relationship-centered approach is the way to understand punishment of all kinds. The "relationship" between punisher and punished here is, after all, *identity*, which seems to be a rather thin, even tautological kind of relationship. It is also the most inescapable. Talk about no-exit. If we must find a way of moving on with one another in the wake of wrongdoing—moving on in a way that shows due respect to the relationship and to the people in it—we will certainly need to find a way of living with *ourselves*, and just as in the other cases forgetting or an easy forgiveness is not always appropriate. What happens when you are the one you have mistreated or failed to

[81] The material in this chapter owes special thanks to Barbara Herman, whose teaching and feedback over the years so deeply inform the work that no citation scheme could do it full justice. Thank you.

respect? When you are the one you feel you cannot go on with in silence, as though nothing happened? Because we are reflective creatures, capable of turning our attention to ourselves, identity for us does not feel thin or tautological. I can make myself, or particular aspects of myself, a distinct object of my own attention, just as I can make someone else its object. But there is something about serious wrongdoing in particular that can cause a kind of coming apart. This is not merely the coming apart of reflective self-engagement, but a rupture analogous to the rupture wrongdoing can cause in our relationships to one another. It can divide us from ourselves, causing a breach that needs healing. We can ask here, as we have in other cases, what if any valuable work the imposition of various forms of deprivations and burdens might do in knitting us back together.

Self-punishment has a complicated place in our thinking about how to live. It has sometimes struck philosophers as being an easier case, morally speaking, because it is the kind of thing we do to ourselves rather than to others, and so seems to avoid certain tricky questions about authority and autonomy.[82] But if self-punishment seems not to be subject to some of the most difficult objections to, e.g., state punishment, it can also seem to lack its point. However

[82] I am thinking particular here of the work of R.A. Duff, whose communicative view of punishment's value informs my own in many ways. Self-punishment, says Duff, "lacks the coercive character that makes imposed punishment morally problematic." Even if one denies, as I do, that all "imposed punishment" has a coercive character, there is still something in this sentiment to appreciate: In cases of intrapersonal punishment, neither the treatment itself nor the understanding that follows from that treatment where it serves punishment's communicative aim, has an external source, and so its "force" does not seem to raise the same kinds of problems. Insofar, in fact, as one shares Duff's view that part of punishment's value or meaning is to "outward[ly]…manifest suffering which gives symbolic expression to the pain of remorse," then self-punishment is best or most essential form of punishment there is: Where state punishment may give "symbolic expression" to a "remorse" that the person punished does not actually feel, self-punishment is an expression of actual remorse. Indeed, Duff claims, "Criminal punishment…should ideally aim to *become* self-punishment; the proper aim of inflicting punishment on a criminal is to persuade her to accept her punishment, to will it for herself, as a penance for her crime." (Duff 1988, p159)

serious the philosophical difficulties about justifying state punishment may be, few doubt that the state has a serious interest in and even responsibility to protect those who live within its jurisdiction from crime and its effects, and to express certain forms of collective disapproval and solidarity.

It can be hard to see, though, what analogous value self-punishment might serve. In the contemporary moment, especially, self-punishment is often held by philosophers and in the popular culture alike to be an unhelpful, even pathological expression of what is sometimes referred to as "toxic" or "primitive" shame.[83] Shame is a self-regarding wrong-reactive attitude, commonly distinguished from *guilt* by its object. While the object of guilt is typically understood to be some particular thing that we have done or said or thought, the object of *shame* is thought to be *ourselves*. To feel guilt is to feel that we have acted in a way that is beneath us, while to feel shame is to feel that we ourselves are lowdown. While guilt is commonly understood (at least in moderation) to be an occasionally appropriate and productive response to one's own wrongdoing, shame is often thought to be superfluous, harmful, or even a sign of moral narcissism.[84] Where the former is thought to motivate a healthy reflectiveness and repair in our relations with others, the latter is thought to turn us myopically toward our *selves,* and often enough to an unproductive and inappropriate self-loathing and self-harm. *Self*-punishment, then, as an expression of shame, can seem the vestige of a primitive or religious past—hair shirts as penance, inflicted in response to some alleged "weakness" of "the flesh," or some more general inability to conform to an inhuman ideal (often enough a sexist, homophobic, cis-sexist ideal). If this is the right view of self-punishment and the attitudes that underlie it, then they present no

---

[83] Bradshaw 2005; Zupancic 1999; Ivy 1993; Nussbaum 2006.

[84] Morrison 1983; and 1997; Nathanson 1987; Broucek 1982.

interesting philosophical problem, but only a therapeutic one: What's the best way to get rid of them? It is only if this is the wrong picture of self-punishment or of the attitudes it allegedly expresses that it would make sense to turn to the question of how those attitudes, when appropriately harbored, may be appropriately expressed, and of whether that expression ever takes punishment's form.

This chapter will have a different shape than the others. Instead of offering a straightforward account of self-punishment and its possible place and value, I will offer an interpretation of the meaning and significance of a relatively obscure passage from Kant's *Religion Within the Boundaries of Mere Reason*.[85] It is a passage in which Kant appears to argue for the rational necessity or at least licensing of belief in a God who punishes the bad, and extends grace to the good. The usefulness of (and perhaps even demand for) such a belief is that it might work to relieve a certain kind of tension internal to the moral life of agents: We have on, on the one hand, the desire for personal happiness, but not just any happiness, *a happiness of which we are* (morally) *worthy*—a happiness that does not come at the expense of the good. And yet we are, on the other hand, intimately familiar with our own personal history of moral failure, so that even when we are lucky enough to find happiness, it can be hard to feel that we deserve it—that it *belongs* to us given the moral compromises involved in its acquisition. To live in the world is to have, at one point or another, profited from wrongdoing: little humiliations inflicted for the sake of popularity, important promises broken in the service of ambition, failures to be honest with people who deserved to know the truth when the price of honesty felt too high. For some of us, whether because we are especially weak or because we were specially tested, there will be the memory of even more profound failures to wrestle with. To be worthy of one's

[85] Kant 1999, 6:72-6:78

happiness, then, requires some way of making up for the failures of one's past—failures one would not now (one hopes) repeat, but cannot now change. This is a deeply felt requirement. We need a way of atoning—of "getting right with God." So Kant here offers us a picture of a God with whom we might, without rational fault, hope to get right. And this God is one who punishes. And this God is one who, if we are good, extends us grace.

The appeal to a punishing and grace-extending God as the resolution to this familiar psychic tension might seem dubious. Even to suggest it as a possible resolution can begin to make the tension itself feel like a dubious one: Perhaps any psychic or metaphysical "problem" that can be solved by appeal to some conception of a punishing God is one we should suspect at the outset of being pathological—a pathology that religion and religious culture itself helps to instill: *You are bad and must atone*. The human desire for absolution is profound, though, and at least as much the cause of religious institutions and belief as their symptom. If religion has sometimes offered irrational or harmful "solutions," that is not to say that the problem is not a real one, that it is not rationally and morally pressing, or that no better resolution is possible. And what's more, Kant's solution—his picture of divine punishment and divine grace—is not, I will argue, as unattractive as it may first appear. It does not, for a start, turn out to be an endorsement of punishment in any standard sense—not even, I will argue, self-punishment. Kant's picture of grace, and even of *God* is one that might, I believe, have some place in a secular ethics.

Having completed the exegetical project, I will turn, then, to the question at hand: Do we need self-punishment? Is there some place for it in a decent human life? Kant, I will argue, has provided us with a basis for thinking that there isn't, and that we don't. What he shows us instead is that while shame, like other forms of what we might call "moral suffering" (e.g. guilt, anxiety about one's own moral status, hardship endured for the sake of doing what's right) is sometimes

appropriate, it is never either appropriately imposed *as*, nor expressed in the intentional imposition of, some burden or deprivation to oneself. All of these form of moral suffering will, in other words, have a place in moral life, and are, in particular, sometimes appropriate responses to wrongdoing—but none will be appropriately imposed *as such*. What we need, instead, to repair the internal rift that wrongdoing sometimes involves is two-fold: First, a sincere commitment, born out in a larger pattern of action, to the principles violated by the wrongdoing; and second, an analogue to forgiveness, concerned not with *guilt* but with *shame*—a forgiveness not for doing, but for *being*. Call this grace.

A final note, before I begin: The issues of atonement and punishing that I set out here to address are framed as being about each agent's relationship to herself, and to her own past. Of course, insofar as the wrongs we commit are wrongs against others, the process of atonement—even insofar as that is a process of resolving something within ourselves—will often involve making amends to others. To focus in such cases on the *intra*personal to the exclusion of the interpersonal, as though one could atone for wronging another without having to include her in that process, may betray a serious misunderstanding of what atonement requires. Still, coming to terms with one's own past is importantly distinct from the process of reckoning with others about it. Coming to terms with past wrongdoing is, after all, the kind of thing we have to do even when our wrongs were victimless, or so terrible (at least in the eyes of the wronged) that interpersonal forgiveness is unlikely, impossible, or even inappropriate. For my purposes here, then, having already dedicated some time to thinking about interpersonal wrongdoing and repair in the last chapter, I will here consider the problem of atonement as though it were a strictly *intrapersonal* matter, though admittedly in most cases (even those I will take up here), the problems of intra- and interpersonal repair in the wake of wrongdoing are inextricably bound up together.

B. Kant's Religion: *Some Background*

In 1748, thirty-five years before Kant published the *Religion*, a merchant ship called the *Greyhound* was caught off the coast of Ireland in a terrible storm. It was the middle of the night, and as the ship filled with water a then 24-year-old John Newton cried out to God for help. It was a moment he experienced as the beginning of his conversion, and for the rest of his life "endeavor[ed] to observe [the anniversary of that] day with humiliation, prayer and praise."[86] The son of a shipmaster and nonconformist Protestant, Newton had embarked on a career working on slave ships. He was a young man who had developed "a pattern of coming very close to death, examining [his] relationship with God, then relapsing into bad habits."[87] This moment, though, marked a turning point for Newton. He worked five more years in the slave trade before leaving to settle in Liverpool, where he taught himself Latin, Greek, theology, and immersed himself in a small church community for several years before being ordained as an Anglican clergyman. In the years that followed, Newton came to be well known in England for two things. The first was his ardent abolitionism. In 1788 (five years now before the publication of the *Religion*) he published a pamphlet called *Thoughts Upon the Slave Trade*, describing with intimate familiarity the conditions on slave ships, and apologizing for "a confession, which," he said, "comes far too late."[88] He worked for the rest of his life as an ally to leaders of the Parliamentary campaign to abolish the slave trade in England, and lived to see that work complete, dying a few months later. Newton's enduring fame, though, has been as a hymnist. He

[86] Newton 1868, p355.

[87] *Ibid.* p21-22

[88] Hochschild 2005, p130-131

wrote many well-known hymns, the most remarkable of which was an account of his conversion experience as a young man on the *Greyhound*—a hymn called "Faith's Review and Expectation", which has since come to be known by its opening lyric, "Amazing Grace".

For most of us, the experience of moral change will be less cinematic than Newton's account of what happened to him on the deck of the *Greyhound*, the moral space traversed over a life will be less magnificently stark than the swing from licentious young slaver to devoted husband and abolitionist; and most of us, too, do not have on our conscience the weight of anything so profound as a hand in the brutal transport of the enslaved. Indeed, there may *be* no weight more profound. Still, the narrative arc serves as a particularly vivid instance of a kind of human life with which we are familiar: A person is born not malicious, perhaps, but nonetheless seeking her own satisfaction without restraint, and as she grows to moral maturity she struggles with the question over and over again of what she is willing to do, to participate in, to tolerate for satisfaction's sake. Often enough she learns the limits by transgressing them, learning what it really is to do, or to participate in, or to tolerate a certain kind of harm exactly by doing, or participating in, or tolerating it. Over time, patterns of moral behavior and reasoning change. Her moral understanding increases, and she comes thereby to see more clearly her own past, and to the need to reckon with it. She is morally improved, but it is by that very improvement that she comes to experience new and deeper forms of moral suffering, including a mature, articulate shame in place of mere paralysis or inchoate unease that may characterize shame in the child or young adult.

The passage of the *Religion* with which we will be concerned is about this person—the changed man, ashamed of his past. It is about, in particular, three problems about the kind of judgment to which the changed man is liable, and, ultimately, what he is entitled to believe about

and hope for *himself*. I will call these three problems, *the problem of imperfection*, *the problem of self-knowledge*, and *the problem of atonement.*

1. The Problem of Imperfection: How, given the moral weakness and imperfection I still sometimes suffer, even having come so far, could I ever be *justly* judged *good* by a being who sees all?

2. The Problem of Self-Knowledge: How can I, who does *not* see all, be sure that I really have changed for the better? That I have not merely hidden the truth from myself?

And finally, the most difficult question in Kant's estimation—

3. The Problem of Atonement: How can I, even if I have changed, justly be judged good given the wrongs of my past—the person I was then? Have I atoned? Is atonement possible?

We need a way of answering these questions, offering some rational support for the changed man's hope that he, a person who has sinned and will no doubt sin again, could ever *justly* be judged *good*—that he could ever make himself worthy of that judgment. Without some basis for that hope, the commitment to morality would be a Sisyphean task, in which we strive for a goal we can never achieve, and at the expense of our own happiness. The stakes of finding some way of answering these questions is thus our very energy for the moral project—the most important project, in Kant's view, and one that requires all the energy we can give it.

To get a better hold of Kant's way of framing the problems and the solution both, it will help to have a little context. *Religion Within the Boundaries of Mere Reason* is a kind of extended discussion of what role there might be for religious concepts and institutions in helping us to make sense of and abide by our moral commitments. On a view like Kant's, where moral worth is entirely a matter of the free exercise of the will in accordance with principles available to all rational agents, it is not clear why either in practice or principle an individual would require anything like a church or religious doctrine in order to be good. Kant does not appeal to religious

doctrine or practice as a way of grounding or attempting to explain what makes it the case that our moral duties are what they are, or that the moral law is authoritative for us.[89] But human life, even human *moral* life, involves more than a set of grounded moral principles. It involves, e.g., the subjective struggle to see clearly through confusion, distraction, and temptation what our moral duties are, and to find ways of abiding by them, even as they conflict with our other aims. In this struggle to understand what we should do and to persevere in doing it, we might well need role models, interlocutors, friends—in short, *each other*—and a shared culture of art and ideas. We may find these forms of support and tools for understanding, which religion sometimes offers, *useful*, not in establishing the foundations of the moral law, but in the practical task of grasping and abiding by it.

In Part II of the *Religion* Kant considers the (rational) use to which we might put certain religious concepts, first and foremost that of conversion. He employs this concept as a way of understanding the familiar, though in some ways puzzling phenomenon of moral transformation. Our personal experience of moral development is by its nature halting and incremental, stretching across the entirety of even the longest and best human lives. Where things go well, we develop over time in terms of both our grasp of what goodness demands, and our ability to live up to those demands. We never achieve perfection, and even progress itself is not perfect. We sometimes revert, forget, fail to learn our lessons, and so on. At the same time, progress, when it comes, can feel sudden. Newton's experience on the deck of the *Greyhound*, spurred by the nearness of death, is one kind of instance, and there are others. Sometimes transgression itself is the spur. We see, suddenly, *what we have done*, and for the first time feel the full weight of morality's authoritative demand, indistinguishable from the feeling of having failed to meet it.

[89] Kant 2012

This is well-trod territory in our cultural mythologies of moral development, religious and otherwise. These moments of realization are sometimes themselves the resolution of a process. Unruly as he was, Newton was clearly a young person who struggled with warring impulses, in a search for some way of resolving them, and his conversion experience seems to have been as much the resolution of that incremental process as the start of a new one. The phenomenon of moral progress, our experience and representations of it, involve this tension between a sense, on the one hand, of a complex, incremental progress, and on the other, of startling breakthrough—including a specially powerful kind of breaking over at the threshold of moral maturity, in which we absorb, in one way or another, a sense the weight of adult responsibility, and of some identification with or taking on of those responsibilities, though we never come to the point of living up to our sense of them in any perfect way.

To see what this comes to requires a basic grasp of Kant's picture of conversion. Though he aims to explain our (imperfect, incremental) experience of moral change in time, he does so by appeal to a certain metaphysics that will be familiar to those with a larger sense of Kant's work. Conversion, he wants us to see, has an underlying logic that admits of no degree, but which is by nature total. There are, as Kant understands it, at bottom only two principles of action—the moral principle, and the principle of self-love. The principle of self-love tells us to pursue our own happiness, making the ultimate reason for undertaking or refraining from a particular act that it does or does not seem to us to serve that purpose. When we act from the principle of self-love, we act on "incentives of inclination," doing what we want, or what would feel good, exactly because we want to. The moral principle, in contrast, tells us to act in accordance with the moral law, on the incentives that *moral* reasons give us to act. The fact that I made a promise can give me a moral reason to act. The fact that some particular course of action

would be "the kind thing to do" may give me some moral reason to do it. The two kinds of reasons are not, of course, always in conflict. Sometimes doing what constitutes keeping my promise is also doing the thing that would give me pleasure. Sometimes morality is silent, having nothing to say for or against doing what would make me happy, and sometimes things go the other way around.

According to Kant, both principles are present in the will of every personal agent, and every personal agent will have both reasons of self-love and moral reasons. What makes the difference between good and evil is which of these two principles is *prior*, acting to constrain the other. For a period in our moral development, even as we begin to grasp and to be moved by moral reasons, the principle of self-love is, for us, prior or "ultimate". We are disposed to pursue our own happiness. This is not to say that our actions necessarily run counter to the moral law. One may be born with a good and careful nature. Still, though, our reason for doing what is right, when we do, will be that it serves what is for us the more fundamental end of our personal satisfaction.

Conversion in Kant's sense occurs when, for a particular person, the principles are reversed, coming to be ordered properly. Now it is the moral law acting as a constraint on what we may and may not do in the name of our own happiness, rather than the principle of self-love acting as a constraint on when we will and will not choose to act in accordance with the moral law. This ordering of principles—this assuming of either a disposition to do what is right (and only within those boundaries to pursue self-interest) or a disposition to do "evil" (to act that is, for the sake of self-love, and to act in accordance with the moral law only insofar as self-love allows)—is an all-or-nothing proposition. This is the general character of our will: either good or evil. It could not be otherwise. There are two principles of action, and often enough they point us

in different directions. In the functional agent, there must be some priority relation between them—some instruction that tells us not only to which we should defer, but how we ought to reason about questions like that. Priority is not a relation that admits of degrees. We are either one kind of person, whose ultimate allegiance is too the good, or another, whose final allegiance is to self-love. To move from being one sort of person to the other is, on this picture, in a very important sense to become morally *a new man*—a point to which I will return.

The dispositions, though, with their totalizing logic, are not constituents of the empirical, or "sensible" world. The sensible world, on Kant's view, consists in things we can take in and cognize by means of physical and mental sensation. Our behavior is in this sense empirical, as is the train of our thoughts. These are things we can sometimes observe and sometimes know about. There are also, though, things whose existence we have reason to posit, but that we can never experience by our senses—the *supersensible*—and our dispositions, or hearts, or are of this second type. The heart is not the constituent of a spacetime. This change, then—conversion in the sense of the reordering of the principles—is of a different kind than the incremental process of moral change and maturation that we experience in and across time. "This is not," as Allen Wood has put it, "an account of the manner in which men, in time, become good," but of the underlying change of heart that makes our change in time possible.[90] [91]

---

[90] Wood 1970, p224.

[91] We should be careful not only to remember that the change of heart does not happen *in* time, but about indexing it to a particular time. It is not necessarily even the *representation* of something that happens in a moment. Whatever John Newton *felt* on the deck of the *Greyhound*, it wasn't his heart, which is not in the realm of the sensible. And while there was certainly something real in that dramatic moment and what it meant for the arc of his life, he writes, when pressed about the subsequent years he spent working on slave ship, "I was [still] greatly deficient in many respects...I cannot consider myself to have been a believer in the full sense of the word, until a considerable time afterwards" (Newton 2003, p84). Perhaps, then, we should be hesitant about indexing this change of heart to any particular moment or event.

How, though, does this atemporal account of conversion as a restructuring of the dispositions or *change of heart* interact with our felt experience of change over time? We might think of Kant here as providing a kind of *metaphysics* of conversion, but we should be careful of the term. John Rawls cautions, correctly I think, against a certain kind of "metaphysical interpretation" of this idea of reason and character as "permanent and timeless."[92] The wrong way of interpreting it, Rawls says, is as describing a kind of "first origin" of our agential experience in time, as though it were another plane of reality.[93] We ought, rather, he says, understand it "from a practical point of view," as providing us with *a way of viewing ourselves* that helps us to make sense of our experience of agential movement—of the experience of *acting* and of *changing as an agent*.[94] The "space", in other words, that the dispositions "inhabit" is not "a different realm, conceived as ontologically separate from the order of nature," but one of rational representation.[95]

So the changed man is one whose commitment to the moral law now conditions his commitment to seeking his own happiness, and one (in our particular example) who has made some progress over time at coming to understand and abide by those commitments in practice. The first problem to arise for the changed man is the *problem of imperfection*.

B. The Three Problems

1. The Problem of Imperfection

[92] Rawls 2000, p301.

[93] *Ibid*.

[94] *Ibid.*

[95] *Ibid.* p307.

So the changed man is one whose commitment to the moral law now conditions his commitment to seeking his own happiness, and one (in our particular example) who has made some progress over time at coming to understand and abide by those commitments in practice. The first problem to arise for the changed man is the *problem of imperfection*. The changed man is still, of course, imperfect, and no one will be more aware of than he himself. However much he has improved, he will notice old patterns of thought and desire, moments of weakness, and so on.

"The solution," says Kant, to the problem of imperfection "rests on the following":

> "According to our mode of estimation, [to us] who are unavoidably restricted to temporal conditions in our conceptions of the relationship of cause and effect, the deed, as a continuous advance *in infinitum* from a defective good to something better, always remains defective, so that we are bound to consider the good as it appears to us, i.e. according to the *deed*, as *at each instant* inadequate to the holy law. But because of the *dispositions* from which it derives and which transcends the senses, we can think of the infinite progression of the good toward conformity to the law as being judged by him who scrutinizes the heart through his pure intellectual intuition) to be perfected whole even with respect to the deed (the life conduct). And so notwithstanding his permanent deficiency, a human being can still expect to be *generally* well-pleasing to God, at whatever point in time his existence be cut short."[96]

The changed man is well-pleasing to God, despite his enduring and indeed ineradicable imperfection, because God sees the disposition, not the deed. To make sense of the idea of good and bad character as a general matter, we needed to represent to ourselves this these state of an underlying disposition to good or to evil—no degrees, and so, now, we have this idea to appeal to in constructing a way of representing to ourselves a fair standard of goodness that we are capable of living up to. We get out of the problem of imperfection by positing a different kind of God-Judge. If the judgment that matters is one that takes the point of view of a cosmic by-

[96] Kant 1999, 6:67.

stander watching a life unfold, or even of a kind of mind- or *will*-reader, assessing in order a long list of maxims acted upon, then the problem remains. But now that we have a way of representing ourselves in terms of an atemporal, rational nature, we can also have the idea of God-Judge who represents us to Godself in the same way—who "looks" to "see" if, as general matter, one's heart was in the right place.[97]

Being good, then, even perfectly good, does not demand perfection of the kind we will strive for in life, though inevitably fail to achieve. Though I framed this problem as one about the kinds of shortcomings we might impute to a person at the end of a long arc of improvement, it is a point, as Kant says, that applies equally to the person less lucky in the time and opportunities for success she is allowed. Suppose, for instance, that John Newton had been struck down by disease or disaster the month before he finally quit his work on the slave ship forever, or the year before he finally got up the understanding and courage to condemn publicly the condemnable practice of slavery. It is only from this point of view, which admits of no degrees, that we make sense of the idea that even the best of us could be judge "well-pleasing to God"—even the best of us allowed immortality, improving forever. From the point of view of the God-Judge Kant describes, this is nothing—as it *must* be nothing.

2. The Problem of Self-Knowledge

Having solved the problem of imperfection in this way, though, we generate, or at least make more difficult, the *problem of self-knowledge*. The problem of self-knowledge, recall, is that the changed man cannot be certain he has really changed. His awareness of the same old

[97] This God will be perceived by a different sort of faculty, one that Kant calls *pure intellectual intuition*—the faculty by which the supersensible is perceived.

patterns of thought and desire, moments of weakness, and so on, which make him worry that he is the same man after all, will also make him worry that he has, perhaps, never really changed. Perhaps he is just fooling everyone. How can he be sure?

The problem as Kant describes is not (on his picture) just a problem about our imperfect capacities of introspection. This may have been the problem before, when we were understanding ourselves as good or bad based in virtue of things like our actions or thoughts—that the changed man, at least in his more honest moments, knew himself (in the relevant sense) all too well—"the judge within him…pronounc[ing] a stern judgment upon himself."[98] He could not imagine being well-pleasing to such a judge. Now, though, there is a different standard—a new kind of God's-eye-view. From this point of view, there is hope for us. But it is a point of view to which we have, even in principle, no direct access. We can't see our own hearts. So though it is true, on this picture, that we may rest assured that God only judges the content of our hearts, we may have little assurance, given our epistemic situation, as to what its contents are. Kant considers and dismisses the notion that "whoever possess as pure a disposition as is required will *feel* [it]," and can trust this feeling.[99] "[O]ne is never," he writes, "more easily deceived than in what promotes a good opinion of oneself."[100] One is never, on this picture, entitled to a dogmatic or doctrinaire assurance on this front.

The changed man finds in the kind epistemic situation Kant describes a limited confidence in her hope that she has changed. He has both an evidential basis for a *cautious* inference that he is no longer the same man. He observes in his behavior and by introspection an

---

[98] *Ibid*. 6:77.

[99] *Ibid.*

[100] *Ibid.*

incremental progress over time toward the pure expression of the good disposition. He will have learned something about his own habits of mind, preferences, and blind spots. This limited confidence will sometimes be shaken, as he reverts to old routines or behaviors, or fails in new ways he never even thought of before. In these moments he will wonder if he is the same person he ever was. Maybe he's just fooling everyone. Maybe he only seemed well-disposed because his goodness wasn't being tested. This is the kind of doubt Kant seems to think is both epistemically warranted, and in some important sense *good* for us. It is, at any rate, the kind of relationship to one's own self-understanding that is most conducive to the forms of incremental progress in which a good life consists.

We needn't consider any question about God or metaphysics to feel the force of Kant's characterization of our position here. Even the best among us are not without limitations when it comes to self-awareness. Our moments of greatest peace and confidence can turn out, in retrospect, to have been folly. What seems like progress can turn out to be just more of the same. The best of us sometime lack a self-confidence that the worst of us often have in spades. An inability to know ourselves entirely—to understand the dispositions that are the spring of our action and reaction—is a feature of the human condition. Kant's characterization of moral progress in the lived world as limited, incremental, and reason to have only a cautious confidence in the deeper state of one's disposition is resonant regardless of one's metaphysical commitments.

3. The Problem of Atonement

This brings us, finally, to the problem of atonement. This "third and apparently greatest difficulty" says Kant, "is as follows:"

> "Whatever [one]'s state in the acquisition of a good disposition, and, indeed, however steadfastly a human being may have persevered in such a disposition in a life conduct conformable to it, *he nevertheless started from evil*, and this is debt which is impossible for him to wipe out. He cannot regard the fact that, after his change of heart, he has not incurred new debts as equivalent to his having paid off the old ones. Nor can he produce, in the future conduct of his life, a surplus over and above what he is under obligation to perform each time; for his duty at each instant is to do all the good in his power. – Moreover, so far as we can judge by our reason's standards of right, this original debt, or at any rate the debt that precedes whatever good a human being may ever do…cannot be erased by somebody else. For it is not *transmissible* liability…but the *most personal* of all liabilities, namely a debt of sin which only the culprit, not the innocent, can bear…."[101]

"[M]oral evil" he continues,

> "brings with it an *infinity* of violations of the law, and hence and *infinity* of guilt…because the evil is in the *disposition* and the maxims in general (in the manner of *universal principles* as contrasted with individual transgressions): consequently, every human being has to expect *infinite* punishment and exclusion from the Kingdom of God."[102]

So the problem is that we "start from evil" and "an infinity of violations" of the moral law, thus accumulating a debt that looks impossible to discharge. We have established we do not on Kant's view "start from evil" in the sense of malevolence, but in the sense that we begin life moved by unchecked inclination, and that inclination never in the course of human life fully comes under the control our subjective commitment to the moral law. And our violations are *infinite* in the sense that they are, so to speak, evils of (transcendent) *being* rather than *doing*. We have already traded in our old idea of God—the one who judges finite actions and imposes commensurately finite forms of punishment or other forms of atonement—a stint in purgatory, a circle in hell. We now have a God who "sees" only the total "evil" and the perfect good that are the underlying spring of action. And so the "debt" we have accumulated is infinite in the straightforward sense that it is not finite—not the kind of thing you measure or trade in. To be corrupt is to be corrupt.

[101] *Ibid.* 6:72, original emphasis.

[102] *Ibid.*

We all, then, accumulate this debt. The case of John Newton is one involving obvious and terrible wrongdoing, but our sources of moral regret—the signals of our origin in evil—can be and typically are much subtler. Some of us were once, to our shame, angry bullies who humiliated others from spite. Some of us were once so timid or concerned to please that we hurt and disrespected those we loved by our dishonesty. Such failures needn't, though, involve harm to another at all, or in any sense that would constitute an interpersonal breach. Perhaps in my youth I made a practice of misrepresenting myself as a friend acting for reasons of love, when in fact I was acting for the sake of social advantage. Maybe I was so good at it that no one noticed, or ran in a crowd of others who were doing the same. That there is no victim, or at least none with whom we have reason to seek repair, does not get us out of the problem, nor does the fact that our sins are of the more mundane variety.

Neither, interestingly, does youth. There is a common story that I take to be a part of this genre of shame-triggering personal history to do with ingratitude toward one's parents in childhood, which typically involves the rejection of a gift or hurling of an insult that one can see in retrospect (and sense in some vague, terrible way at the time) is humiliating or hurtful to the parent. I have a story like that, and probably you do, too. Jeffrie Murphy tells his own in the wonderful paper, "Shame Creeps Through Guilt and Feels like Retribution".[103] It is such a

---

[103] "When I was a boy, I loved baseball and desperately wanted a baseball glove. Although I was not aware of this, my parents were at that time experiencing great financial distress – such great distress, indeed, that even the cost of a baseball glove was beyond their means. As it happened, the naval base where my father was stationed fielded several amateur baseball teams and these teams sometimes discarded used equipment. My father came across a discarded base- ball glove. He tried it on, found that it was still in surprisingly good condition, and – with delight in his eyes – presented it to me when he came home at the end of the day. But my father was, alas, left-handed; and I am right-handed – a difference he had not, at the moment of trying on the glove, brought to consciousness. When I tried it on, my immediate disappointment at its uselessness to me was obvious and I rather contemptuously cast it aside. My father lost his temper, called me

common genre that I even have a favorite.[104] It begins early, then, this sense of a debt. And it is not the sense of a debt owed to other person, but one can exist in the absence of any victim, or a morally mature and responsible self. Indeed, such memories may be so searing precisely because they wake us up for the first time to the reality of our moral situation—that we have the awesome power to shame and humiliate others (even those we perceive as being very strong, or whom we love very much) and the awesome responsibility that comes with it. And yet these incidents often lay heavily on the conscience.[105]

---

ungrateful and selfish, and sent me to my room – behavior on his part that I now see as his way of defending himself against his own hurt and disappointment when one of his rare attempts to do something he found so difficult – show love – misfired" (Murphy 1999, p338-339).

[104] My favorite is recounted by the writer and comedian Rob Delaney. He recalls, in his memoir, his exhausted, single, working mother making him a birthday cake with a frosting-art portrait of all four members of his favorite band, Danzig. "It was a great cake," he writes, "and, when she presented it to me, I became infuriated":

> "'Mom! Come on! Danzig shouldn't be on a cake! They're like, bad dudes! They would never be on a cake! Maybe they'd be, like on a tombstone or a gunslinger's coat, but a cake! No way! Jeez!'"

"I want to cry," now he says, "thinking about the pain that was on her face."

> "Here was a woman who worked full-time at an insurance agency, working hard to support her two kids, always making sure to be extremely present in our lives—mornings, evenings, and weekends. She had worked on her masterpiece in secret, studying each band member's scowl, to make her self-proclaimed bad-ass little boy an extremely cool cake, and he hated it. And he let her *know,* like a real piece of shit."

"To this day," he writes, "when I imagine having a time machine, my FIRST stop is my thirteenth birthday where I would jump up and down with excitement and hug my mom when she reveals that cake. If there was still time left on my time travel visa, only then would I go back and kill Hitler." (Delaney 2013, p22-23)

[105] "[I]t is one of the more interesting things about human psychology," says Murphy, "that such small incidents can become life-defining moments for us. They can, though generating no guilt or shame or bad conscience at the time, generate all this to a painful degree when recalled in later life." His own memory is one, he says, "that I recall with enough painful guilt to say with confidence that I feel in it the deep pangs of a bad conscience" (Murphy 1999, p339).

This debt the changed man has is one that he is stuck with. It remains, Kant says, even if he stops adding to it, it cannot be paid by another, and the good he does now does not itself generate a "surplus" he could use to pay it down. Ultimately, Kant *will* appeal to some idea of grace, and some idea of paying one's moral debt by good acts as the solution to the problem of atonement, but he rejects them here in these forms: First, that a living savior without sin could bear the debt of sin in our place. It is too "personal" a liability, he says, to be transmissible. This is not, it seems to me, a moral point so much as a logical one: If I punish the innocent for the sins of the guilt, then the sin has not really been punished at all.

The sense in which we generate no moral surplus by our good acts is also grasped easily enough, and born out by our own sense of things. Consider John Newton, dedicating the final years of his life to the cause of abolition. While there is much to admire in his rare dedication to the cause of justice, it is dedication which he himself felt (correctly I think) was no more than what morality asked of him—not as a former slaver, but as a human being. His dedication may have been rare, but that does not *per se* make it supererogatory. It seems to be a common feature of the very good that however extraordinary their acts, they take themselves to be doing no more than what the world calls them to do. Regardless of whether one thinks that they are right, we are trying to solve a problem that arises (even) from their point of view, and in that context, appeals to an alleged surplus won't do.

The solution Kant suggests emerges in three steps. The first echoes the solution to the problem of imperfection: "The judicial verdict of the one who knows the heart of the accused must be thought as based on the universal disposition of the latter, not in the appearances of his

dispositions."[106] We add, though, in this section, a new thought—or, at any rate, a new point of emphasis—that "after his conversion…he leads a new life and has become a 'new man'."[107] "Conversion," Kant continues, "is an exit from evil and an entry into goodness, '*the putting off of the old man, and the putting on of the new*', since the subject dies unto sin in order to live unto justice."[108] From the God's eye point of view, then, the man who sinned is simply gone—transformed—and the man before God's judgment has nothing to atone for—is not the proper object of "punishment [or] exclusion."

Someone has to answer for those sins. "[S]atisfaction," Kant says, "must be rendered to Supreme Justice, in whose sight no one deserving of punishment can go unpunished."[109] So now we have a puzzle: The old man deserved punishment but never got it—indeed, he was well compensated for his sins. The new man doesn't deserve it, but he is all that's left. The solution to this particular problem—the only logical move left—is that "the punishment must be thought as adequately executed in the situation of conversion itself."[110] This is the second step. "We must therefore," he says

> "see whether, by means of the very concept of moral conversion, we can think of that situation as entailing such ills as the new human being, whose disposition is good, can regard as having been incurred by himself (in a different context) and, [therefore], as *punishment* whereby satisfaction is rendered to divine justice."[111]

---

[106] Kant 1999, 6:72.

[107] *Ibid.*

[108] *Ibid*, emphasis added.

[109] *Ibid.*

[110] *Ibid.*

[111] *Ibid.* 6:74, original emphasis.

Is, in other words, any of what happens in the process of conversion the kind of thing that we might conceive of as a punishment, which the changed man could regard as being for the sins of his past? He has a particular candidate in mind: the "long train of life's ills…which the new man undertakes…for the sake of the good."[112] These ills will include "all the sufferings and ills of life in general," and in particular the kinds of hardship and self-sacrifice that come with subjugating one's own self-interest to the moral law—of doing without things that cannot be justly attained, being the bigger person when provoked, and all the challenges we face in seeing this organizing commitment through.[113] This train of ills is not, for the changed man, a punishment—either divine of self-inflicted—but an incidental feature of an undertaking that's just about doing what's right. And yet, if those hardships are not a punishment for the changed man (and would not be fitting as such), they "are still fitting *punishment* for someone else, namely the old human being (who, morally, is another human being)."[114]

Kant here gives the changed man a new way of thinking about this incidental feature of his life—the difficulties inherent in doing right by himself and others. He can think of them as burdens that, though for him are *not* a punishment, *would* be for the person he once was. And this hardship that he willingly bears is one he can think of as having this second moral aspect: that it pays down the moral debt that the former self incurred, thus rendering "satisfaction" to "Supreme Justice."

So if the first move in answering the problem of atonement is to say, roughly, "you are no longer the same man in God's eyes—the debt is not yours", the second is to say "and yet you do

---

[112] *Ibid.*

[113] *Ibid.* 6:75.

[114] *Ibid*. 6:74.

pay it—you are paying it"—not by generating excess good, but by bearing costs in the course of conversion that you otherwise would not have. To see how this works requires that we return for a moment to the relationship between the two dimensions of the change of heart. While the noumenal change of heart is total, the change of heart as experience by the changed man is one that unfolds incrementally over the course of his life as an agent, and so his whole life is, in fact, a slow death and rebirth, and the rebirth, as it happens, will always involve a new willingness to take on hardship or sacrifice, if necessary, in order to do what is right. It is in this way that the hardship born across the span of his "new" life is something we can rightly think of as a cost born in transition from old to new.

So the changed man can be entitled, on this view, to think that there is reason to hope the person who did those things is not him, and that in the work of becoming new, he bears a cost that, were he to see his experience of conversion through to its full completion, would constitute a full reckoning and clean slate for the person he would then be. But we do not, of course, on this view ever see our experience of conversion through, but are "always only in mere becoming."[115] And this is why we need a third move, the crucial move, which gives us a way of making up "the surplus over the merit from works."[116]

This third move in solving the problem of atonement is to appeal to an idea of grace—one like and unlike the one with which we are familiar. Grace, recall, is God's unearned favor, secured for us by the sacrifice of another—a being without sin, who bears the cost for us of our own. Kant's notion of grace is also a kind of unearned favor, by which we think of God as imputing to us a goodness we have only achieved in part *as if* we already possessed it here in

[115] *Ibid*. 6:75

[116] *Ibid.*

full"—"And to this we indeed have no rightful claim."[117] We need a God who is willing, if we do our part to make up the difference. Grace—the imputation of perfection—is "unearned" in the sense that it is *not owed.* It is not possible, even in principle, for the human being to do all the work that would be required to make it the case that she is owed, as a matter almost of justice, God's pleasure. Even with her best efforts, she will still be in need of grace. But what she *can* do by her best efforts—and what she can only do that way—is make herself *worthy* of God's generosity. She does this by taking upon *herself* the burdens of goodness, rather than by belief in another who allegedly bears those burdens on her behalf.

And yet the idea of a vicarious substitute is, in slightly altered form, still important on this picture of grace and redemption. Consider Kant's language here:

> "...this [good] disposition which he has incorporated in all its purity, like unto the purity of the Son of God—or (if we personify this idea) this very Son of God—bears as *vicarious substitute* the debt of sin for him, and also for all who believe (practically) in him: as *savior*, he satisfies the highest justice through suffering and death, and, as *advocate*, he makes it possible for them to hope that they will appear justified before their judge. Only we must remember that (in this way of imagining) the suffering which the new man must endure while dying to the *old* human being throughout his life is depicted in the representative of the human kind as a death suffered once and for all."[118]

The "vicarious substitute", then, is not a historical figure—a fellow spacetime inhabitant—who takes the debt of sin for us. The vicarious substitute is, rather, a way of *representing* our own suffering back to ourselves in a helpful way. The reality on this picture is that the changed man is *his own* vicarious substitute. He is two—the new man, and the old; the sinner who never pays, and the innocent who willingly bears the cost of a debt of sin. In the story of the crucifixion we have the idea of "a death suffered once and for all," which makes vivid for us the demand that *we* (still mired in the disposition to evil) suffer a death—a death that is also a rebirth—but which

---

[117] *Ibid*.

[118] *Ibid*. 6:74.

will (if everything goes well) stretch across the whole of our lives. We save ourselves, by making ourselves worthy of approval in the only way available to us: By our devotion to a principle, demonstrated in our doing as much as we can with as long as we have.

C. Representing God to Ourselves

The passage from Kant with which we have been concerned represents itself as being about how it is that we, imperfect as we are, could be entitled to the hope of being well-pleasing to God. But we might just as well think of it as a passage about what kind of *God* we would have to believe in to be entitled to that hope. In the course of the passage Kant refers several times to other conceptions of God that he dismisses as inadequate. From these asides I discern four distinct (though sometimes overlapping) conceptions of what God might be like:

The first conception of God Kant considers and dismisses is what I will call the Cosmic Jurist—a judge with perfect fact-finding capabilities, and an infallible sense of proportionality. The Cosmic Jurist knows everything about your life that you know, and more besides. He sees your life pass before His eyes, and hands down a verdict and a sentence. Because, Kant says, we can only assess our *own* moral status by examining our actions and other forms of sensible evidence, we are "not able to think of any other condition of being delivered to the verdict of a future judge…than that [*our*] *whole* [*lives*] be one day placed before the judge's eyes."[119]

The Cosmic Jurist has some advantages. Namely, Kant says, that the idea of a God-Judge who knows everything that *we* know about ourselves, so that we can't pull the wool over His

[119] *Ibid*. 6:77.

eyes, keeps us properly anxious about our moral status. While Kant doesn't say more here about the shortcomings of the Cosmic Jurist, there are things we can infer from the differences between it and the conception of the God-Judge Kant advances. The Cosmic Jurist only knows the kinds of things that *we* know. He knows what we do, what we think, but his judgment of who we *are* will be, like our own, just a decent inference. We should be looking, instead, for a God who sees *who we are*, where this is something more or other than the sum of our actions and mental lives. It is, after all, the anxiety about who we are—about what our worst thoughts and actions might mean about us—that makes us worry. That God will execute a perfectly informed and uncompromising judgment on this matter is what makes God an object of our highest hopes for, and worst fears about ourselves.

Kant also briefly dismisses the concept of what I will call the Venal God-King, whose approval we can buy with flattery or ritualistic offerings. This is the conception of God for whom Kant reserves his greatest contempt—the God who can be "mollif[ied]…with prayers and entreaties…with incantations and self-proclaimed expressions of faith."[120] The Venal God-King is essentially useless as a means of assuaging our need for absolution. He is a God we can distract—neither able nor concerned to know our hearts—and whose judgment need not reflect what he *does* know, as he can be swayed and manipulated because He has His own needs—greed and venality—to assuage. The Venal God-King, though, can offer us no meaningful redemption. The terms of salvation He offers bear no relevant connection to the deepest fears and anxieties Kant claims we harbor about ourselves, which are not that we are incapable of bribery or flattery, but that we have fallen short in moral terms. Neither the condemnation nor the "redemption" offered by this kind of God means in anything in those terms. The problem that the Venal God-

[120] *Ibid*.

King solves is, rather, the problem that arises for the "evil" or unchanged man, whose highest principle is still self-love. For this person, perhaps, it is enough to engage in such a transaction, paying in gold or flattery for the promise of a comfortable afterlife, or to assuage guilt and shame, understood as merely uncomfortable feelings, to be avoided as such.

The third conception of God that Kant implicitly and explicitly rejects in this passage is the Vindictive God, with a desire for blood. This is a God for whom "penitential…expiations" from the whip or the hair shirt can make any difference to one's moral standing. We appease this judge by "remorseful self-inflicted torments."[121] I will say more momentarily about Kant's relationship to this idea of suffering as inherently redemptive, but for now it is enough to note that Kant mentions this idea of God and rejects it in favor of one who demands that we "impose" upon ourselves a burdensome righteousness, and not just that we suffer any burden—let alone a violent one.

There is, finally, a forth conception of God to consider—one with which we are familiar, and find traces of, at least, in Kant's language. Call this the God of Love. The God of Love, like all but the Cosmic Jurist, is not a kind of projection of the "judge within" us, but "another judge, of whom news [of us] will be had from other sources of information elsewhere."[122] The God of Love sees us at a distance, as part of a species, understanding and accepting our "human frailty," and granting on that basis a kind of blanket amnesty, or grace without condition. The God of Love might be disappointed, but He isn't angry.

Why reject the God of Love? This is a God who does much of the same important work that God on Kant's conception does, without having to appeal either to a retributive

[121] *Ibid*.

[122] *Ibid*.

understanding of punishment or to a standard of human conduct that demands a perfection which is out of reach, and then tells us we need to be forgiven for our failure to reach it. The conception of God appeals to the (perhaps correct) notion that our native imperfections, delicate constitutions, and contingent fortunes make holding-accountable inappropriate, at least outside of the context of our personal relationships. The God of Love also seems to be effective in precisely the way we care about, relieving certain feelings of shame and anxiety that turn us inward, away from the humble, outward-facing project of helping our neighbors, and so on.

The God that Kant constructs for us is not any of these—certainly not the God of Love. Kant's God measures us against a standard that even the best of us will never be able to meet. That the standard of conduct is not sensitive to human frailty, though, does not mean that Kant's conception of God is insensitive to it. Indeed, his is a God "designed" exactly to put salvation within reach, and designed, too, to give us relief from those same feelings of shame and anxiety, and for just the same kind of reason. What this conception of God does *not* do is treat these forms of rational relief and self-acceptance as available to everyone, regardless of what we are like. It makes these conditional, demanding something for them.

Perhaps there are good reasons to think that this is a genuine downside of this picture—that a God who makes salvation conditional seems less *good* to us, and so less God-like. Or perhaps it seems to assume a picture of desert that strikes us as itself irrational, and so not appropriately built into our conception of the highest arbiter. Even if we were to grant this for the sake of argument, though, there is an important trade-off to consider: When we lower the bar, we make the "prize" less attractive, perhaps even bleeding it of its essential value—not because God's approval is only valuable to us if others are excluded from it, but because it comes so cheaply, demanding literally nothing of us. It is a picture on which the young slaver is as entitled

to satisfaction he finds in a life of terrible misconduct at the old abolitionist is to the satisfaction he finds in a life of service and advocacy, and the old abolitionist as being just as entitled to the satisfaction he finds in a life of service and advocacy as he would have been had he never been a young slaver. None of these characters, on the God of Love conception, is more or less entitled to happiness, nor more or less entitled to relief from their respective experiences of shame. And the view I take Kant to hold is that such a conception of God should and will fail to satisfy not some inhuman standard, but *us*. It will fail to satisfy the young slaver and the old abolitionist most of all. It attempts to relieve the anxiety we feel at being unworthy of approval and of happiness by answering that there is no such thing as unworthiness. In demanding nothing of us, the God of Love is rendered little better than a drug for forgetting. We need a God who sets terms we can meet, not one who sets no terms at all.

E. The Place of Punishment and Other Forms of Self-Imposed Suffering

What, finally, is Kant's picture here of punishment—or, at least, of deserved suffering? There is a range of hardships that he holds to be appropriate in the life of the changed man. The first is moral anxiety. Kant seems to hold that as a descriptive matter, the changed man just will harbor a certain amount of uncertainty about his own underlying character, which will be a source of discomfort for him. He also seems to hold that this anxiety is, in some measure, both warranted and useful: warranted because our own underlying character is not something we *can* know directly, even by careful introspection, and about which we are entitled to make only cautious inference on the basis of long experience. The more recent one's conversion, the more limited the evidence, the more such anxiety is warranted. This anxiety can also, Kant speculates,

be productive, serving as a way of keeping us both humble and vigilant with respect to the state of our own character in a way that aids in our efforts to improve incrementally over time. (I do no take Kant here to be suggesting that this form of moral anxiety is warranted *because* it is useful, nor would I find such a position defensible.)

While moral anxiety is, on this view, a form of moral suffering appropriate to the situation of the changed man, it is not a form of "deserved suffering", let alone a form of punishment. It is not, first of all, a response, even in some metaphorical sense, to any wrongdoing on the part of the agent. One might say he is in this position because he "started from evil", but on the picture Kant has given us, this would be no more than to say that he is in this position because he is human. It is not, further, a form of suffering intentionally imposed on the changed man—either self-imposed, or imposed by anyone else. It is no more or less than the situation in which we find ourselves, which is one that involves a kind of uncertainty that the changed man, who cares deeply about norms and values he aims to live by, will of course feel the weight of. To fail to acknowledge that we *do* feel this weight, and that there is no (rational) way out of it, is not to adopt a condemnatory view of human nature, but just to acknowledge and articulate one of the inescapable burdens of moral agency.

The next candidate form of moral suffering is the hardship and self-sacrifice involved in doing the right thing. This would seem to be our best candidate for being a form of punishment, as Kant describes it in these terms: but not a punishment for *the changed man himself*. It is not a punishment, recall, *exactly because* the reasons one has for "imposing" the hardship upon oneself are not to do with the value of the hardship. The changed man undertakes the right course of action not *for hardship's sake*, but *despite* hardship, for the sake of the independent right-

making reasons he has for doing it. I tell you the truth, despite my strong disinclination (given the cost to myself it will involve) *because you deserve to hear it from me*.

If part of my reason for doing the right thing is *because* I am disinclined to, though, and because I think it would be good for me (given my own current or former bad character) to suffer the discomfort of doing the good thing I would prefer not to do, then it would, indeed, start to look like a form of self-punishment. When Kant speaks of the need for the changed man to *welcome* the hardship and self-sacrifice involved in doing right, it can start to feel that he is, indeed, heading down this path—or, at least, as though the changed man will look forward to the opportunity to show off his feats of moral strength. This would be, I think, an unwelcome outcome, and there is good reason to think that Kant would agree. One of the right-making features of a situation then would seem to be *that it would allow me to demonstrate my righteousness*. But Kant's is not a view that seems to license, let alone congratulate those who do what's right for that kind of reason. I am inclined, then, to offer a friendlier reading, on which the changed man is not one who welcomes hardship and self-sacrifice as a demonstration of moral devotion, but one who finds meaning, and therefore strength, in the hardship and self-sacrifice morality demands, thus transforming what would otherwise be the experience of imposed suffering, to suffering undertaken in the name of one's own principles and values. No punishment here.

But what, finally, about shame and its expression? Surely the shame we feel at who we once were is something Kant holds is sometimes appropriate. It is the initial premise of the conversation that human beings harbor this shame at our past moral weakness and failure, and that there are appropriate and inappropriate ways of relieving oneself of that shame. It is not, in all cases at least, a mere nuisance, of which we might appropriately rid ourselves by taking a pill.

The happiness we seek in its relief is only something that the changed man *wants* insofar as he can see his way to believing that he deserves it. It is not wrong, then, to say that on this view shame (and also guilt) is, where appropriate (which it sometimes is), a form of deserved suffering. This does not seem to me, though, to be a special feature of Kant's moral view in particular. It is the same conclusion anyone, even the non-retributivist, must reach if they think that self-directed wrong-reactive attitudes like guilt and shame are ever warranted, and warranted on the basis not of utility, but as a response to one's own moral failures. And this experience of shame is not, anyway, a form of punishment for exactly this reason: Shame—at least appropriate shame—is not something we impose upon ourselves (or have imposed upon us) *as* a burden. It is a feeling that is burdensome, but not one we impose as such—or one that we *impose* at all, anymore than we can be said to "impose" sadness upon ourselves, or disappointment.[123] Shame itself is no punishment.

What, though, about punishing expressions of shame? Surely these are sometimes forms of self-punishment—indeed, the very forms with which we were initially concerned. Here, though, we find that Kant thoroughly shares the modern distaste for self-imposed punishment as either an expression of shame or an attempt to relieve a sense of that feeling. He has dismissed—*contemptuously* dismissed—the idea that self-imposed suffering redeems. Such suffering, his idea seems to be, is hollow, *useless* suffering. It opposes the suffering endured for the sake of being and doing better—which is, after all, what endlessly matters. Although he may have lacked the language to say it, Kant's arguments jibe with the spirit that animates contemporary thought

[123] I do not claim that we can never be said to impose feelings like sadness upon ourselves, but only that such feelings will not be warranted. This would be to, e.g., feel sad at the loss of a friend not directly because one has lost a friend, but because one feels one *ought to be* sad at the loss of a friend.

on the subject of such forms of self-inflicted suffering. Across the board, we find that that they are not just useless, but narcissistic. Forms of self-inflicted punishment mark a turning away from the actual path to redemption, which involves a kind of selflessness, and in, *toward* the self, as though one's suffering might, in itself, be any more redemptive than one's personal experience of pleasure. In this way, self-flagellation can, on Kant's view (and on the common understanding), turn out to be nothing more than a perverse form of self-love.

The only "expression" of shame that seems supported on this view is the channeling of that regret into moral self-improvement. But even this is not quite right. It is not the case that I should tell my friend a hard truth that she deserves to hear from me as a way of expressing my shame at past failures to do so. It is, rather, that when shame flares up, I can (a) take a limited comfort in the thought that I don't seem to be that truth-reticent person anymore, and (b) turn away from that feeling of shame at past reticence, insofar as it has the power to arrest and paralyze. In this way, I can engage in the only form of action that could make me worthy of that feeling's relief.

F. The Take Away

What value could any of this have for a contemporary, secular ethics? And how, in particular, might it inform a relationship-centered account of self-punishment? When Kant sets out to answer the question of what God would have to be like in order for the changed man to be, as he hopes, well-pleasing to God, he sets out to answer the question of what kind of God we can turn to in the hope of a meaningful forgiveness. This forgiveness is a making-whole or restoration of the self, in the way that forgiveness between persons can constitutes a resolution in or restoration of a relationship. The problem Kant sets himself to resolving is what I have called

the problem of atonement, which is, quite literally, the problem of *making one thing of two*.[124] It is the problem of alleviating alienation by making amends. While Kant frames this as an atoning to God, it is a conception of God better understood as a way of thinking about our relationship to ourselves. We are looking for a God who represents to us in a more perfect form both the moral and rational commitments that structure our own point(s) of view, and the sense of ourselves as hopelessly imperfect realizers of those commitments. We need a way of reconciling these within ourselves, and the God we might have some rational license in which/whom to believe, or to employ as a heuristic, will be one that gives us a satisfying way of doing this. What we need is not a way of becoming one with God, but of achieving unity within ourselves.

When we do serious wrong—violating moral norms for the sake of serving our own interests or desires—we thereby alienate ourselves *from* ourselves, and one shape that alienation can take is shame. Shame is, in short, an attitude of alienation from ourselves. It takes the self, either now or in the past, as an object of alienation and even revulsion. Sometimes these attitudes are straightforwardly misplaced. Instances occur in which we ourselves have not done wrong, yet we experience shame nevertheless. We may, for example, have simply been mistreated by another, and that mistreatment occasioned in us a sense of self-rejection or self-directed revulsion. In shame I reject a person, who is *me*, as someone with whom I either cannot identify or do not wish to, because they (and yet I) treated carelessly what I care for deeply, defiled what seems to me sacred, or failed to demonstrate a virtue I exalt. My caring and my valuing and my respect for those things are essential parts of who I take myself to be. I want to believe that offending and offensive person isn't me. There is, perhaps, even a sense in which I *must* believe

[124] "Atonement" or "the condition of being at one" derives from the compound "at-one-ment", where the verb *to one* was understood as meaning to *make* or *put at* one. (see "Atone" and "Atonement," *OED* 1971, p539).

it—believe that I, as I know myself to be, would never do such a thing. To believe *of* yourself, *about* yourself that you would never defile whatever it is you value seems to be integral to what it *is* to value something deeply. That is to say that to take yourself to be a valuer (of whatever) is to believe that you are not (wholly) yourself when you fail to value it. This is not to say that we never really act in violation of our values, or even that we are incapable of predicting that we will, but just to say that when we do, it will generate a kind of psychic tension that finds its relief in some distancing: Who am I? Am I the person who values x? Or am I the person who *violated* that value? To answer the former question in the affirmative, we need a way of answering the latter question in the negative.

At the same time, though, I know that I *am* in an important sense that person. Sometimes this will be in the straightforward sense that I haven't changed, or worry I haven't. I still seem to be a person claiming a committed to x, but likely to fail in that commitment. I am not yet in a position to claim with any authority that the violator was someone importantly different from the person I am now. I am, at best, in the position simply of wishing I was.

But even if I have changed—if friends and foes alike agree that I am a different person now—I still have a problem. I may, of course, still have private doubts about myself, but even putting those aside, there is something else. I may not be that person anymore, but I live in her house. I have her job. I raise her children; I am loved and supported by her parents. I am her heir and inheritor of her ill-gotten gains. We know, for example, that John Newton, a slave ship captain turned staunch abolitionist, quit the slave trade and denounced it outright, but the life he began as (by all accounts) a good husband and clergyman was undertaken with resources secured by his participation in human trafficking. For the whole of his life, his material wealth was thus entangled, and it, along with many a business and personal relationship, would remain so until he

died. Even if, after his change of heart, he had given away every literal penny he made in naval wages, severed every tie he forged as a merchant, and vowed never again to participate in commerce bolstered by the slave trade, Newton's past moral commitments would *themselves* remain part of this inheritance. His abhorrence for the slave trade—an abhorrence so deeply felt and movingly expressed that it helped to change the shape of the world forever—was born of his participation in it. Newton's past wrongdoings proved a benefit that he not only *couldn't* give back, but could not possibly even *want* to. The majority of what we gain by and from wrongdoing cannot be wiped away. There is no way of giving back, either in principle or on pain of violating the terms of some other important commitment. There is no help for it. We must continue on in the lives we have. So we need a way of making them ours—something to which we are *entitled*, so that we can feel at home in our own lives, and, having paid off that debt, let the old self go.

Kant gives us in his picture of conversion and of the God-Judge, a way of representing to ourselves both the distance and the intimacy that characterize the relationship between a person and her shameful past—a way of cutting loose the past on fair terms. It is not an escape from responsibility. It involves harboring no illusions about ourselves. And yet it points a path toward letting go—not for the sake of personal satisfaction, but for goodness' sake.

In the course of offering solutions to these three problems, it seems to me that Kant offers at least two suggestions that should be interesting, especially to philosophers thinking about the proper place of shame in moral life. The first is in the conceptualization of the hardship and sacrifice involved in moral life as meaningful in this particular way—as a kind of *gift*, freely offered, "from the 'new man' on behalf of the 'old'—"an interpretive device," says Barbara

Herman, which adds a genuinely "new piece to our moral repertoire."[125] We must take care, I have cautioned, in how this device is employed. I am cautious, myself, of Kant's notion of "welcoming" hardship as a "test" or (as we contemplate the prospect of what sorts of hardships we should be willing to take on) as a debt-paying mechanism (even if the debt is someone else's). Thinking in this way, surely, can risk the kinds of dark and objectionably self-regarding thinking about moral life that we (and Kant) should be anxious not to endorse. But it can also be a tool that we employ in exactly the moment when shame catches us up and threatens to paralyze. When shame says, "You deserve to suffer," threatening to pull us into patterns of self-harm, it also gives us a reply: "I do, but there's no need to put aside my work to go looking for suffering. Suffering will come in the course of things." This knowledge of suffering's imminence, its status as life's undercurrent and occasional overflow in retrospect or even as kind of shading over our lived experience of hardship, can render it especially meaningful for us strikes me as being a productive source of comfort—one without any cost that I can see. For all the talk of Supreme Justice, it is a view that ultimately offers sensible and not unkind advice: Get to work. Try hard. Be good. Let the rest take care of itself. And when the past catches you up so that you're afraid it will derail you, here's a way you can think about it that might help.

The second interesting feature, and a closely related one, is Kant's decision to employ the concept of grace. Grace belongs to a family of concepts, not least mercy and forgiveness. The logic of these concepts is in contrast (though not necessarily in tension) with notions of justice and right. The latter category of concepts includes those things that we can owe to others, and those that they may rightfully demand of us. The former, by contrast, are human goods that cannot be owed or rightfully demanded, yet they operate on the logic of gifts or blessings. These

[125] Herman *forthcoming*, p16.

goods are often associated with the supererogatory, and with virtues or imperfect duties of kindness and generosity. And yet, on the other hand, they are all things of which we seem to be able to make ourselves more or less worthy. But while mercy and forgiveness have a well-delineated place in our secular ethics, grace does not. We do sometimes speak of "showing" one another grace, by which we mean a kind of generous overlooking of one another's human imperfections. But even this tends to be a secondary use of the term, which is mostly (in my limited experience) employed by those who also understand and employ it in its primary, religious sense. I know of only one (unpublished) attempt in the philosophical literature to work out a secular account of grace and its place in moral life.[126]

Grace, though, like forgiveness and mercy, seems to me to be a concept that can float free of its specifically religious meaning, while retaining its distinctive shape. "Unearned favor" is something anyone might extend, and that we might gratefully receive from a variety of sources. This unearned favor would look (as in the religious context) like being accepted or judged "well-pleasing" despite our flaws—flaws that others would be well within their *rights* to complain about. We extend grace to one another in this sense all the time, and it might be interesting (as with forgiveness and mercy) to think about what makes it the case that this extending sometimes is and sometimes is not appropriate, along with a range of other analogous issues. Here we might think of Kant as offering the reasonable though by no means uncontroversial suggestion that its appropriateness will depend on the general character of grace's potential object. (Is it in some sense *fundamentally good*, despite its flaws? Is it moving in the right direction?) This concept

[126] In her dissertation, and in an as-yet-unpublished paper, "Grace: Goodness in Love in the Bad", Vida Yao talks about grace as *love not fully explained by the excellence of its object*, and attempts to work out a place for this secular notion of grace in thinking through some conceptual problems about the relationship between (appropriate) love and goodness, especially in the context of personal relationships between agents.

may be especially important in thinking about the ends and ideals of particular kinds of personal relationships. I am struck, for example, by how much certain clinical descriptions of what constitutes a happy marriage sound like living in a state of mutually extended grace.

But it seems to me that we should be especially interested in thinking about the potential (and even potentially necessary) role that some concept of grace might play in helping us to understand what an ideal relationship to *self* might look like. This is a complicated question. An ideal of the relationship we stand in to ourselves seems to fold in aspects of many others. It is incumbent upon us sometimes to *parent* ourselves, sometimes to be a *friend* to ourselves, and as Kant's text suggests, even the way we understand our relationship with *God* might be understood as a way of modeling something about the relationship we ought to stand in toward ourselves. But it seems likely (and in keeping with the relevant aspects of these other forms of relationships) that this ideal will involve some kind of (limited) self-understanding and (conditional) self-acceptance. Kant's idea on this front may be helpful to us in thinking about what it looks like to be self-regarding things, both committed to the good and doomed to fail at it all the time and all over the place, in need of an attitude to take toward ourselves that can fully accommodate both features.

A relationship-centered account of self-punishment begins exactly here: establishing an account of the relevant form of personal relationship. First, we would need to think that we stand in some meaningful moral relationship to ourselves, and, second, we ought to establish what the central aim or value of that relationship might be. It seems to me undeniable that there is a morally important identity relationship. Without one there could not, at minimum, be such a

thing as duties to oneself as a person.[127] What the import of that relationship might be, and what its aims are, may vary from moral theory to moral theory, but what Kant's work here suggests is a view on which the aim of the relationship we stand in to ourselves is to be good and to do right. Any adequate account of self-punishment would then speak to the ways in which punishment of some particular variety could work to advance that aim.

Kant, it seems to me, takes seriously the idea, which is at the heart of a relationship-centered account of punishment, of wrongdoing as first and foremost constituting or causing a kind of rupture. Subsequently, he speaks to the necessity of finding some means of repair in the wake of such rupture. In other contexts, I have argued that some reactions to wrongdoing that will be at least occasionally crucial to serving this aim are going to constitute a form of punishment. But, on my interpretation, Kant offers an account of how reconstitution in the wake of rupture might involve nothing that we would consider punishment. His work suggests that the forms of burden and deprivation that *will* be valuable to us for such purposes will not be *imposed* burdenings or deprivings at all. They will simply occur in the natural course of things or, at least, not be *intentionally* imposed, but rather incidental features of some course of action taken on for the sake of another, better purpose.

It makes sense that incidental burdens and deprivations should be sufficient for repair in the intrapersonal case, but not for the interpersonal one. Recall that in the case of friendship, the aim of punishment (where it is permissible) is to bring it about that the wrongdoer *understands what she has done.* In the intrapersonal case, by contrast, the rupture that punishment might permissibly work to resolve is one that only arises where the wrongdoer *already* understands

[127] There are those who doubt that there is such a thing as duties to self (see Singer 1959, or Williams 1985, p202).

what she has done. It is only once she understands herself to have engaged in serious wrongdoing, and has committed (or recommitted) herself to the moral principles she thereby violated, that she is divided from her past, and so in the market for some way of reconciling herself to that past without compromise. It is the John Newton who had come to understand his past complicity in the slave trade as a terrible wrong who felt the need of some way of getting right.

Chapter 5: The Relationship-Centered Theory of Punishment and the State

A. Introduction

In Chapter 1, I began by denying three claims: that (1) punishment cannot occur, or occur permissibly outside of legal and institutional contexts; that (2) our standard theories of legal punishment can be easily extended to characterize or provide us with proper conditions on its permissibility; and that (3) punishment in interpersonal contexts has nothing of central import to teach us about the nature and possible legitimacy of state punishment. Over the course of the dissertation I have begun to make a case for denying both (1) and (2). I have argued that punishment is a wider phenomenon than the traditional definitions have admitted—broad enough even to capture cases that we might not initially consider in those terms. I have attempted to defend some such instances of punishment as permissible. I have defended these cases not by appeal to standard theories of punishment, which begin from an account of punishment's general justifying aim, but by constructing a series of novel accounts of punishment's justifying aim in the context of a particular relationship, each account grounded, in turn, in some account of the nature of the relationship in which the relevant form of punishing occurs.

The aim of this chapter is two-fold. First, I will step back and offer a very general initial sketch of the relationship-centered approach to theorizing punishment, as against the standard alternatives. While the justifications I have offered have differed in crucial ways from one context to another, they represent a unified and distinctive approach to the subject of punishment, which can be applied across the full range of possible contexts in which the question of punishment's potential place and value might arise. This approach to theorizing punishment constitutes an alternative to the major theories on offer. It is distinct both from so-called "backward-looking" retributive approaches on the one hand, and "forward-looking" deterrence-type approaches on the other. It puts repair front and center—the repair, in particular, of the relationship between punisher and punished. Where a friend, or a neighbor, or a colleague, or a citizen violates the terms of that relationship in a sufficiently serious way, the relationship sustains damage, and there will be cases in which the other party to that relationship can contribute meaningfully and even crucially to its repair by responding in a way that constitutes an intentional deprivation or burden. It is these cases, in which the punisher acts for the sake of the relationship as an independent source of value, that we should be concerned to defend.

Second, I will return to the question of state punishment in particular, suggesting that we need a relationship-centered account and offering some very initial thoughts about the general form that a relationship-centered account of state punishment might take. What I have to say here will not itself amount to a full relationship-centered account of state punishment, but a sketch of how we might begin to approach the subject of state punishment. The primary question I will be concerned with here is; *what kind of relationship* should a relationship-centered account of criminal punishment be concerned with? Here I will also spend a fair amount of time considering existing theories—this time theories already in the neighborhood of the kind of relationship-

centered approach I have in mind.

I consider first the restorative justice approach, noting that while it is a kind of relationship-centered approach to state punishment, it is a relationship-centered approach that centers on personal relationships, even in the context of the public, criminal law. A relationship-centered account of the criminal law, I argue, should instead attend most directly to the relationship between the wrongdoer and the *state*, where this is conceived of as a meaningful kind of relationship in itself—a relationship that criminal wrongdoing damages, and that criminal punishment might work to repair.

Next I will consider Joshua Kleinfeld's *reconstructivist* approach to theorizing criminal law and punishment. Reconstructivism begins promisingly, placing repair and the value of a certain kind of public, political relationship front and center. It does not, though, characterize the central relationship that criminal punishment aims to repair as being the relationship between the state and the *offender*, nor pursue the question of how punishment might work to achieve that repair, bringing offenders back into good standing. On Kleinfeld's picture, the wrongdoer seems to recede in importance, operating merely as, first, by his wrongdoing, a threat to the social fabric, and then, by the suffering of punishment, as the means whereby citizens already in good standing work to repair their political relationships to *one another* by demonstrating their shared commitment to the norms of conduct that the wrongdoer has violated.

The right kind of relationship-centered account of punishment is one that puts the relationship between punisher and punished front and center, and takes the legitimate aim of punishment to be the repair of *that* relationship. In the case of criminal punishment by the state, that relationship is the one that holds between the state and the offender. Once we establish that this is the relationship to which we should be attending there is, of course, much remaining work

to be done in order to deliver even a rough sketch of a proper relationship-centered account of punishment in this domain. I will end, then, by offering just an initial sketch of what that work might be. It will involve, at minimum, establishing something more concrete about what an ideal of the relationship between citizen and state might be; how, precisely, certain violations of the law work to damage that relationship, and what those violations are; and how, precisely, certain forms of punishment might work to repair that damage, and what those particular forms of punishment might be. While I am not yet prepared to answer these questions, I will end by laying them out. In a couple of cases, I will offer initial thoughts about the sorts of answers that seem more or less promising.

B. The Standard Approaches: Forward- and Backward-Looking Theories of Punishment

One way the literature on punishment has traditionally been organized is around a distinction between so-called "backward"- and "forward"-looking accounts of punishment. Backward-looking accounts justify punishment as a response to the past, to settle old scores. These are typically accounts on which wrongdoers are said on the basis of past bad behavior to deserve hard treatment, and on which the state is thus, under certain conditions at least, entitled to impose it.[128] On forward-looking accounts, by contrast, punishment aims to bring about a certain kind of future—one in which there will be less wrongdoing and so less suffering and

[128] The view that punishment is a form of *score-settling* or repayment and the view that it is a way of giving wrongdoers what they deserve are not, as Cottingham reminds us, identical, but two logically distinct (though often closely related) forms of retributivism (Cottingham 1979, p239). Both, though, are (as are all purely retributive theories of punishment) instances of backward-looking theories. Cottingham also reminds us that merely to have "a backward-looking element in a theory" is not sufficient to make it a distinctively *retributive* theory (though he points out that this distinction is largely ignored by many philosophers).

greater security than there otherwise would have been.[129] Such views generally focus on the role of punishment in deterring crime. [130]

Both forward- and backward-looking theories of punishment alike have been sources of chronic dissatisfaction. Critics of backward-looking, retributive justifications of punishment have found the central notions of desert and score-settling to be hopelessly mysterious, if not flatly immoral. Such critics may be skeptical of the idea that people can, as a general matter, deserve to suffer,[131] that a person can deserve to suffer some *particular* punishment for any particular crime,[132] or that—even if we can makes sense of this notion of desert—the state is entitled to go around giving people what they deserve on that basis.[133] These objections about the nature of desert are instances of a more general kind of worry. "Desert" names the reason we are said to be

---

[129] The best example of a purely forward-looking account of punishment is Jeremy Bentham's. On Bentham's view "all punishment in itself is evil," and "if it ought at all to be admitted, it ought only to be admitted in as far as it promises to exclude some greater evil"—where "greater evil" means some future state of affairs involving even greater suffer (relative the future state of affairs that would obtain were we *not* to punish) than the suffering the punishment itself will bring about (Bentham 1789, chXIII.2). This kind of purely forward-looking justification was also popular among mid-twentieth century criminologists (see Wootton 1963 and Menninger 1968).

[130] While there have been theories of punishment that that we describe in other terms, they tend nonetheless to fall pretty easily into one of these categories—the one exception being restorative justice theories, which I will discuss at length before the end of the chapter. Consider, for instance, punishment as understood on moral education, or paternalistic theories of punishment. Some moral education theories are ultimately pretty standard hybrid theories, on which the general justifying aim of (state) punishment is to deter crime, but to do so in such a way as to be consistent with respect for persons, the state must only punish inside certain constraints (i.e. by aiming to deter criminal behavior *in a particular way*—by appealing to wrongdoer as a rational moral agent). (See Hampton 1984). Other moral education theorists, like Nozick, offer what look like pretty standard retributive theories, but argue that in giving the wrongdoer what he deserves, we benefit him in some way (Nozick 1981). Likewise expressive views, which do not posit punishment's alleged expressive function as an alterative to deterrence or retributive aims, but as a *means* by which we deter, or as an additional function

[131] Scanlon 2008.

[132] Shafer-Landau 1996.

[133] Dolinko 1991.

entitled to impose a certain kind of harm, regardless of whether or not that harm serves as the instrument of any further good or purpose. Punishment's value, on such theories, is said to be *non*-instrumental. But for those who feel that the imposition of hard-treatment demands some further justification (especially when the hard treatment in question constitutes a form of state coercion), appeals to desert seem almost question-begging. To such critics the imposition of hard-treatment seems to be a kind of prima facie wrong that had better at the very least be good for something.

If, however, we adopt a theory on which punishment is justified for the sake of some further good, we run into a different set of problem. Perhaps we are merely using those we punish (often in a particularly harsh and violating way), treating their right and interests as something we are entitled to trade away for the sake some further or "greater" good. Critics of forward-looking justifications argue that punishment as conceived on such theories fails to treat agents with due respect.[134] On this type of view punishment is of primarily instrumental value, and is permissibly done *only* for some further purpose (e.g. sending a message to others). Even rehabilitative theories, on which punishment allegedly works to serve the good of the offender, the primary justification for the practice of punishment is one of public safety, where this end is said to justify substantial interference in the lives of those who have committed some criminal offense. So while on the one hand we think that punishment, if it is justified, must be *for* something of real value, we also worry, on the other, that punishment (especially, again, when the hard treatment in question constitutes a form of state coercion) might amount to an objectionable form of using the *object* of punishment—the wrongdoer—for something.

If we conceive of the forward-looking/backward-looking distinction as carving up and

[134] See especially Hegel 1991, Morris 1968, Murphy 1973, and Harrison 1988.

exhausting the logical space of justification, there will seem to be only two alternatives: Either join the ranks of those who say "both" and attempt to defend a hybrid theory, or join the ranks of those who say "neither" and hold that punishment is never, even in principle, justified. Neither option seems fully to settle the problem. Hybrid theories aim to capture the virtues of both approaches, demonstrating that state systems of punishment (or idealized versions of these systems, at any rate) seem to have both retributive *and* deterrence aspects, and helping us to sort out the level at or degree to which the different theories operate or act as constraints on one another.[135] Such theories also, though, inherit both sets of problems. They do not redeem retributivism, which they endorse at some level or in some degree, to those who hold that persons cannot deserve to suffer, or for whom the very notion of desert is hopelessly mysterious. Neither do they redeem consequentialist accounts of punishment's justification, to which they also help themselves, to those who hold that respect for persons is inconsistent with the practice of harming or interfering with the autonomy of one person for the sake of others to whom she constitutes no immediate, comparable threat.[136]

It seems, then, that in terms of assuming a general orientation toward the question of punishment and its justification one is left either to pick one's poison, or to become a skeptic, rejecting even the possibility of permissible punishment. This has seemed to many philosophers an appealing—or, at least, dangerously enticing—route. We speak of punishment as a practice that raises very difficult, even intractable philosophical problems, so that even if there are

---

[135] Rawls 1955, Hart 1968, Scheid 1997, Primoratz 1999.

[136] There are, of course, better and worse attempts. Mitchell Berman has offered his own challenge to the traditional distinction between retributive and deterrence-type theories, a good overview of the problems with traditional hybrid theories, and attempts a better integrated dualist approach—one that does not marginalize retributivism in the way that that Rawls and Hart did (Berman 2008).

relatively few pure skeptics, we treat skepticism as an ever-looming possibility. "[T]o listen to philosophers discussing [punishment]," said Jean Hampton, "one would think [it] impossible to justify and difficult even to understand."[137] And yet, as Hampton also notes, "there are few social practices more time-honored or more widely accepted throughout the world."[138] The bare fact that punishment has been widely practiced throughout human history is not, of course, sufficient reason to think that there must be a satisfying justification for it. It would not be the first time that such arguments have been offered in defense of some irredeemable form of human subjugation. Still, we should be awfully careful about offering categorical condemnations of a practice so basic to human life—of concluding that it could have no possible value or place there, however unlike its current place that may be. Especially if one accepts that punishment can take the much broader and subtler range of forms I have suggested.

It might be that the quest for a more satisfying justification is indeed best pursued from inside one of these four approaches. Perhaps there is some answer to the perennial objections to either forward- or backward-looking theories that we have not yet reached. Perhaps we must learn to accept that skepticism, while practically onerous or even tragically demanding, is morally correct. Perhaps the question itself is so difficult or the practice so fraught, that we should not expect even the right answer to be fully satisfying.

C. An Alternative: The Relationship-Centered Approach

But when all possible answers to a question seem wrong, one ought to at least consider the possibility that the question itself is the problem. Perhaps it is a mistake to think that

[137] Hampton 1984, p208.

[138] *Ibid.*

punishment is ultimately about the future, the past, or some combination of the two, and a mistake to organize our thinking about punishment around that distinction. Punishment, I argue, is neither "forward-looking" nor "backward-looking", but *relationship-centered.* Punishment is *about*, that is, neither the future nor the past, but about the relationship between the punisher and the punished, and the harm done to that relationship when one party wrongs the other in a sufficiently serious way.

Over the course of the last two chapters I have offered accounts of punishment in several domains of human life, which should be understood as neither forward- nor backward-looking, but relationship-centered. From these particular accounts, a broader relationship-centered strategy for theorizing punishment begins to emerge. Adopting this strategy will involve understanding the arc of wrongdoing and redress in something like the following terms: Wrongdoing disrupts the norm-governed equilibrium between two agents (i.e. their relationship). Punishment, where permissible, aims to restore that equilibrium (or, in the case of childhood punishment, to teach children how to go about the fraught and essential work of restoration). Insofar as punishment is permissible at all, then, it is an answer to the question, *What now? How do we go on together, given what has happened?* It is not a matter of looking to the past, or the future, or to both, but about "looking" *to the relationship*, to the harm it has sustained, and also moving past it, where these are not two things, but one: An attempt to repair damage done.

It is very easy to confuse the significance of a relationship and the damage it may sustain in wrongdoing with its temporal directions. After all, a relationship is the kind of thing that has (at least from some point in time) a future and a past. Relationships, like individuals, unfold across time, a fact that structures our experience of wrongdoing and redress, as it structures our experience of everything. It is understandable, then, if the "direction" punishment "looks" is

relationship-ward—and, more particularly, toward the damage a relationship has sustained—that we should notice its significance with respect to the future and the past, and debate the place of each in an adequate justification.

Still, to think that punishment could be about the future or the past or about some combination of the two is a subtle but fundamental kind of error. Punishment is a response to wrongdoing. When we have violated the terms of a relationship, thus wronging the other person in it, there is a sense in which that wrong was done in the past. We can, in many cases, name the time and date and location at which we experienced the wrong's occurrence. In another sense, though, the harm is not stuck in the past. If you have betrayed me, and the damage to our relationship (by your violation of its governing norms) has gone unrepaired, then the problem is in an important sense about the current state of our relationship.[139] The problem isn't trapped in the amber of history. Nor is it simply a problem about securing goods for the future. Something of value to us—both instrumental and non-instrumental—has been damaged, and to value it in the way that we do is to bear some responsibility for its care and maintenance.

While relationship-centered accounts of punishment are neither forward-looking, deterrence-type theories of punishment nor backward-looking retributive ones, such theories have the potential to capture what is right in both while avoiding the fundamental stumbling blocks of each. A relationship-centered account of punishment is, like deterrence theories, an instrumental account, on which punishment is not a good in itself, but works to serve some purpose or greater value (or must, at least, in order to be permissible). Legitimate punishment must be *for* something—something we can understand. At the same time, though, the legitimizing value that it serves will not merely be the "greater good", nor the interests of the

[139] Hieronymi 2001.

punisher, but the good of the relationship between punisher and punished, which is itself an independent source of value. Offering a relationship-centered account of punishment thus allows for the possibility of an instrumental account, on which punishment is only justified insofar as it serves some further value, but without running into the traditional problems of instrumental theories. Punishment, that is, may serve some further purpose (where that purpose is repair of the relationship), without thereby treating her interests as something to be traded off for the interests of others if the price is right, or her rational capacities as something to be manipulated for the purpose of serving others' ends if those ends are sufficiently important.

A relationship-centered account of punishment not only provides a way of taking seriously the retributivist's concern about respect for persons (both victim and wrongdoer), but of getting at the grain of truth in the notion, central to retributivism, of moral debt. In spelling out what desert comes to, retributivists tend to reach for metaphors of rebalancing, often monetary. Such metaphors are, on the one hand, just that—metaphors. They are of limited use in telling us with any real clarity what punishment's value *is*. On the other hand, they may suggest a misleadingly simple and objectionably transactional picture of what that value might be. Still, the notion of punishment as extracted payment of a moral debt grips us, and does so because it gets at something of fundamental import. There is indeed something here, of which we can make sense: It is the state of equilibrium that is the dynamic between agents in good standing with one another—an equilibrium that wrongdoing disturbs, and that punishment, I will argue, insofar as it serves any value at all, aims to restore. Call this dynamic, sometimes in equilibrium sometimes not, a relationship.

The general approach to theorizing punishment that has emerged over the following several chapters takes something like the following form:

- Punishment's *aim* is reparative, and it works to serve a *collective* good—the *relationship* (which is a source of value to its members, collectively).

- Its *mechanisms* by which punishment works to serve this aim are communicative (and, occasionally, also work to *teach* communication).

- The *authority* to punish is just the authority one has in virtue of being a member (sometimes a particular member) of the damaged relationship—or, at any rate, someone responsible for the well-being of the relationship.

- The *permission* to punish is grounded in and constrained by the norms governing that particular kind of relationship.

What, then, can we say about criminal punishment in particular? For the remainder of this chapter, I will focus on one small piece of the puzzle: Namely, *what the relevant relationship in an adequate theory of state punishment might be.* In order even to get clear about what we should take the a proper aim of a relationship-centered account of state punishment to be, we have to know what the relationship is, and this, it turns out, is itself a controversial question, even among those who agree that the justifying aim of punishment is the role it plays in repair.

D. Two (More or Less) Relationship-Centered Views

Consider two theories of punishment that we might well think of as relationship-centering alternatives to the standard backward- and forward-looking theories of punishment: The first is restorative justice models. Restorative justice models of punishment offer one kind of relationship-centered account of punishment's legitimate aim.[140] On restorative justice models, the legitimate aim of a criminal justice system is to facilitate reconciliation between the offender and his victims, and to a lesser extend the offender and the broader community in which the offense occurred. The state here is meant to intervene in ways that aim to provide offenders with

[140] See especially Braithwaite 1999, Cragg 1992, Johnstone 2007, and (for a more retributivist-flavored restorative justice model) Bennett 2008.

the opportunity to redeem themselves, by, e.g., acknowledging and taking meaningful responsibility for their wrongs, and making restitution to victims where possible. Restorative justice model acknowledge that it is not only the relationship between the offender and his victim that will be impaired by criminal wrongdoing, but between the criminal and his extended community. Community members are enmeshed in a network of relationships. If I violate a neighbor's right against being physically assaulted, I may not only impair my relationship with him, but with his friends and family, who are also affected. I may impair my relationships with other neighbors, to whom I now seem to constitute a threat, insofar as I am not someone they can trust not to violate their most basic rights.[141] The restorative justice approach begins by putting this relationship-impairing feature of criminal wrongdoing front and center, and demanding that we think of punishment's value in terms of how it may or may not constitute a path toward repairing those relationships.[142]

---

[141] Restorative justice practices tend to handle the impairment of this wider set of relationships not by treating the entire community or some subset of it (e.g. the victim's family) as *themselves* victims, but by inviting them into the criminal justice process in a variety of other ways. "In RJ crime is seen as harming the victim, the community, and the offender, thereby also setting off a spider web of effects to others that puts things out of balance…All of the interested parties," then, "are included in attempting to resolve the harm and restore balance" (CERA 2015). Members of the broader community may be brought in (especially in juvenile cases) as part of a *family group conferencing* practice, which brings in a network of friends and family for both victim and offender. Non-familial community members may be brought in as part of a *restorative conferencing* practice (Morris 2001).

[142] This may actually understate the restorative justice theorist's concern with personal relationships, insofar as such theorists often understand criminal wrongdoing not only as itself relationship-impairing, but as a reliable indication that the relevant relationship was already impaired. Braithwaite writes:

> "Part of the restorative justice theoretical perspective is that disputes will rarely or never be lacking in important implications for human relationships and will often have their source in problems with human relationships. There is an essential claim here that human beings are relational animals…Restorative justice…is about widening the agenda of legal disputes to relational rifts that might be healed" (Braithwaite 2001, p243).

Some restorative justice theorists (as well as many critics and explicators of the view) have characterized it as a form of abolitionism, according to which the state ought not punish at all. Restorative justice theorists put front and center, e.g., ritualized testimony of victims before their wrongdoers, and of subsequent apology. They emphasize the place of making restitution through service as an alternative to, e.g., incarceration. In this way "[r]estorative justice," claims Joshua Kleinfeld, "in its radical form aspires to create a new social order with gentler norms…[It] is committed to the proposition that a world without hard treatment is possible and would be functional." On this way of understanding the view, it says that insofar as what we ought to be aiming for in the wake of wrongdoing is the repair of relationships, we ought not punish at all, but focus rather on more constructive interventions, which provide a way back *into* community for wrongdoers, rather than further alienating them from it.

This characterization of restorative justice models—even "radical" ones—is misguided. The sorts of responses to criminal wrongdoing such models counsel might well constitute punishments on the broader view of punishment I have offered. As in the case of childhood punishment, where we sometimes disagree about whether some particular way of responding to wrongdoing constitutes an alternative to punishment or an alternative *form* of punishing, it helps to begin by clarifying what we mean by punishment. On the view I have defended in Chapter One, "gentler" interventions may still constitute hard-treatment. Jean Hampton makes a version of this case herself, arguing that the real character of "hard-treatment" is not that it involves pain or suffering, per se, but that it constitutes a "disruption of the freedom to pursue the satisfaction of one's desires," which is something that may occur when, for example, a wrongdoer is required

to engage in community service, or perhaps even, she suggests, endure a lecture.[143] While there is room to debate what forms of deprivation do and do not potentially constitute hard treatment, we should be careful about prejudging the cases.

There is, though, good reason to think of at least some restorative justice theories as abolitionist with respect to the state practice of *criminal* punishment in particular, not because the interventions they suggest don't constitute punishments, but because punishment in these cases is characterized as being on behalf of private citizens and their personal relationships. This (relatively dominant) strand of the view emphasizes that we should not conceive of the state as a wronged party, but focus instead on the persons who have been victimized or otherwise effected by criminal wrongdoing. On this kind of view the state's role in punishing is, if anything, to play host to a set of proceedings in which victims and community members do the demanding and the imposing, or where, at least, it is done on their behalf. On this view it is the actual persons in that community that have been meaningfully wronged, but not the state itself. Further, it is only on behalf of those actual persons that state punishment is permissibly imposed.

This position represents a sort of relationship-based account. If *A* is only permitted to punish *B* in response to some violation on *B's* part of the terms of their relationship, and if one holds that the law-breaker's violation is an essentially *personal* one—a violation of the terms of his relationship with the victim and members of their community—then it is they who ought to be doing the punishing, or on whose behalf, at any rate, the punishment must be done. Because the state is not considered the wronged party by restorative theorists, the wrongdoing is, in some sense, none of the state's business, expect perhaps insofar as it might usefully act as referee or

[143] Hampton 1984, p224.

representative (for the sake of ensuring fairness, say) in what is at bottom an interpersonal matter.

We need not disagree with the restorative justice theorist who feels that there is a harm done to the victim and to the community by certain kinds of law breaking, and that these personal harms (or some personal dimension thereof) should not and even cannot be addressed by an agent of the state. Incarceration imposed by the state may be as useless for the purposes of helping heal the relationship between a thief and the long-time neighbor he stole from, and we can appeal to the relationship-centered theory of punishment to say why. When we look to the norms governing that particular relationship, we should expect to find things like mutual respect for basic rights, among them certain kinds of property rights, a willingness to provide aid or stand in solidarity when providing it would be an easy matter, and so forth. That a neighbor who has violated these terms has served time in jail, paid a fine to the state, or even completed a course of community service does little or nothing by way of repairing the harm done to the personal relationship he has to his neighbor. It is simply beside the point.

Where we need not, and should not agree with the restorative justice model, though, is in thinking that there is not *also* some role for a state practice of criminal punishment, where that practice works to repair not the personal relationships between citizens, but the relationship between citizens and the state. A relationship-centered account of punishment that ignores or denies the possibility of this form of *institutional* relationship or its central place in a theory of *criminal punishment in particular* is missing something crucial. As an explanatory matter, such an account leaves it mysterious why we draw a distinction between public and private and law, treating criminal wrongdoing as a violation of public trust and not merely the terms of our private relationships. We *have* a sphere of private law in which individual citizens who have

suffered some unfair harm loss in violation of a legal obligation on the part of the wrongdoer can seek redress or not as they see fit. We also have a sphere of criminal law and criminal sanction, in which the state or "the people" prosecute wrongdoing in violation of the law. We seem to think that there is something important about *prosecuting* crime as a matter of public justice, where this is distinct from merely settling the private grievances of the citizens who have been directly or indirectly victimized. Public prosecutors may work with the victims of the crimes they prosecute, and, one hopes, exercise their prosecutorial discretion in a way that is sensitive to their wishes. But in the sphere of public law it is not the victim's right to decide whether or not the state will proceed in a criminal prosecution.[144] The state, we seem to think, is entitled and sometimes even obligated to prosecute those who violated the legal rights of others against harm, even where the victim is indifferent or even opposed to the prospect. In a relationship-centered theory of punishment that only acknowledges personal relationships, it would turn out to be an odd and objectionable feature of the law.

The restorative justice theorist may consider this a diagnosis rather than a theoretical difficulty, arguing that a better, more defensible system of criminal law would be reconfigured to center personal relationships and the rights of victims. What would be missing, though, in the world that the restorative justice model imagines is a concrete, public manifestation of our commitment to the value of the lives of all citizens. Consider a case in which a serious form of assault is committed against some member of a small, rural religious community, by a stranger, say, and in which both the individual victim of that crime and the larger community prefer as a matter of their personal faith-based commitments to wave their rights against the offender. Such

[144] For an extended discussion of the rights of victims in criminal prosecution, and of the legal and moral responsibilities that prosecutors have toward crime victims, see Goldstein 1982 and Gershman 2005.

commitments may be worthy of admiration, but a criminal law under which offenders were not prosecuted for their assaults on the lives of, e.g., the especially forgiving, would run afoul of the very idea that should serve as its basis: that all lives are valuable, and that one citizen's right to, e.g., bodily integrity is the same as another's. The restorative justice model fails to take seriously the idea that there are *im*personal norms—both rights and obligations we have in virtue of our relationships to social institutions. To take this idea seriously is to recognize this relationship as one that wrongdoing can damage (in addition to whatever damage that wrongdoing may wreak in the context of our personal relationships), and this damage as something that certain forms of punishment might work to repair, insofar as (specifically criminal) punishment amounts to a concrete, public manifestation of the state's obligation to take seriously the equal value of citizens' lives and basic liberties.

This brief sketch, of course, leaves open a variety of important questions. We would need to start from a fuller account of the relationship between the state and those who live under its laws—the constitutive aims of that relationship, and of the right and obligations it involves. From here we would need to know something about how criminal wrongdoing does or does not constitutes some harm to the state in that context, or damage to the relationship itself, and, finally, what if anything punishment in some of its myriad forms might do to repair that damage. We might also wonder how, for instance, such an account would handle white-collar crime, or what such an account might have to say about the authority of an illegitimate or oppressive state to punish. What matters for now, though, is just to say that we should be looking for the kind of relationship-centered account that centers the relationship between offenders and the public institutions *whose norms his criminal offense constitutes a violation of*.

In his 2016 paper "Reconstructivism: The Place of Criminal Law in Ethical Life" Joshua Kleinfeld offers such an account.[145] Reconstructivism is the view that punishment's central, justifying aim is the reconstruction of shared moral norms, as codified in a system of criminal law, in the wake of their violation. Abstract moral norms are on this view given their force by the place they hold in the "embodied ethical life" of a community. When members of the community violate or "break" some abstract moral norm, the damage is done not to the abstract norm itself, but to the community's commitment to that norm as "embodied" in the shared moral life of that community. When we write an abstract moral norm into binding law, this is a way of "embodying" the norm. When individual community members take that law to be binding, and act according to their commitment to the law as codification of the abstract moral principle, this is a way of "embodying" that principle. And when the community refuses to tolerate violations of that law by other members of the community, this, too, is a way—a crucial way—of embodying the norm. If the collective and individual actions of community members in accordance with an abstract moral norm work to build, constitute, and reinforce that norm in the community's embodied ethical life, then its violation by members of the community will work to destabilize and degrade it.

Reconstructivism does a better job than restorative justice models at helping us to focus our attention on a fundamentally public and political form of relationship, and thereby treating the criminal justice system as an enterprise concerned with certain institutionally mediated relationships rather than personal ones. But while it gets something fundamentally right in its characterization of the role of the *punisher* in legitimate state punishment, it misses something

[145] Kleinfeld picks out as part of this traditional philosophical sociologists like Durkheim and Weber, and philosophers so such as Hegel, Jean Hampton, and Herbert Morris.

crucial in how it characterizes the role of the *punished.* The restorative justice models can sometimes seem to treat the state's role in punishment as merely instrumental in what is ultimately the resolution of a personal matters. Reconstructivism, on the other hand, treats the *offender* as an instrument, used to serve the wider good of solidarity amongst non-offenders.

According to reconstructivism, the ultimate value of punishment lies not in giving wrongdoers what they deserve, nor in maximizing happiness, but in *securing solidarity* among citizens.[146] It is solidarity, not happiness or desert that serves as the "lodestar" of the theory: that good which punishment (insofar as it is worth defending) works to serve. In stabilizing and reinscribing the shared moral norms of a community—the norms governing conduct among them—punishment serves to unify them under terms that make shared life possible. To be in solidarity here is a kind of political relationship, involving an understanding of others are *worthy* of the forms of respect that the law demands we show them. We are said to be in solidarity with other members of our community insofar as we understand ourselves and others to be morally required to, e.g., refrain from acting in ways that would constitute a violations of their basic liberties. Solidarity is something that we are, by its nature, in *with others*, and which has a public nature. It is not something we *have* but something we are said to *show* or to *stand in*.

According to reconstructivism, both the construction and maintenance of a community's embodied moral life and its degradation are accomplished through the expressive power of human activity. Some activity expresses our commitment to the abstract moral norms in question, while others express our contempt for the same. When individual members of a society violate a norm of conduct to which the broader community has demonstrated some commitment, they express contempt for that norm, and insofar as that norm presupposes or aims to protect the

[146] Kleinfeld 2016, p1492.

equal moral status of community members, their violation of the norm will also reflect their sense of the relatively degraded status of the victim. To victimize a member of one's community is to express both contempt for a set of abstract moral principles, and for the victim himself, whom the wrongdoer failed to treat as someone with equal standing, deserving of recognition as such. To punish, on this view, is to rebut the claim that wrongdoing constitutes—it is a denial of the claim that the violated moral norm is not worthy of respect, and that the victim is of lesser moral or social standing.[147] To fail to punish is to let the claim of the wrongdoer stand, and to allow this claim to be woven into the fabric of our embodied ethical life, thus degrading the force of our shared commitment to the abstract moral norms and their force in our lives.

Reconstructivism offers a genuine and compelling alternative to retributive, deterrence, or hybrid views. The value at its heart is not mysterious or morally suspect, but one that is entirely respectable, whatever one's meta-ethical commitments. And reconstructivism, unlike deterrence theories, does not license the instrumentalization of the punished in at least one

[147] This brief gloss elides, for the sake of brevity, two distinct claims Kleinfeld makes. The first claim is that punishment is a way of denying the truth of wrongdoers' claims—of saying "this claim about the abstract norms is false—they are worthy of respect." The second claim is that punishment is a way of *making it the case* that the claim *is* false. This second claim relies on Kleinfeld's notion that the dignity and respect owed to persons, like any set of abstract norms, has a lived instantiation that gives it force. In the case of this particular set of abstract norms, he calls this lived instantiation "social dignity" (as opposed to "*ultimate human dignity*, that quasi-mystical core of worth that human beings are thought o have in virtue of being human") (Kleinfeld 2016, p1508, original emphasis). One can be said to have social dignity only to the extent that the other members of her community *treat* (and otherwise acknowledge) her as a person worthy of basic forms of respect. One of the claims that the wrongdoer allegedly makes in victimizing another member of the community is that the victim lack's this status—that the members of her community *do not* take her to be the kind of thing they are required by their shared moral norms to treat as worthy of these basic forms of respect. When the community itself then punishes the wrongdoer, they don't merely rebut the wrongdoer's claim in the weaker sense of offering a counter claim, but in the stronger sense of actually *making it the case* that the claim is false (*Ibid*. p1508-1509).

important sense: it does not treat the interests of the punished as something that can be traded off for the sake of others'.

For those moved by these standard objections to retributive and deterrence theories, reconstructivism provides an appealing alternative. What Kleinfeld fails to do in his presentation and defense of reconstructivism is offer arguments that should convince the skeptic. Or, put another way, Kleinfeld has not yet provided a sufficient answer to the permissibility question: "But why are we entitled to serve that (admittedly valuable) aim in *that* way?" One may well agree that solidarity is surely an important, even central value, agree that punishment can work in service of that value, agree that it is the special responsibility of a particular person or set of persons to act in service of that value in such cases, while still holding that punishment is not a morally permissible means of serving that aim.

Kleinfeld responds to the skeptical worry in several ways. First, he makes a move that will be familiar to the reader by now, arguing that there is no other way of achieving punishment's aim except by punishing. "In the final analysis," says Kleinfeld,

> "reconstructivism does not need to explain why our social language is the way it is. Regardless of the answer to that question, reconstructivism submits as an empirical contention that if our society were to let criminal offenders go in favor of the alternatives the say-it-with-flowers objection suggests, our norms would be insecure. If we tsk-tsked and didn't punish, our norms would not be reaffirmed in an effective way by other means; they would simply collapse." (1523)

Punishment, he asserts, just is our convention for expressing solidarity in the wake of certain, sufficiently serious forms of wrongdoing—that no other response does, as a matter of fact, express that which solidarity (and so, ultimately, human thriving) requires us to find some way of expressing.

Kleinfeld's second form of response suggests that the philosopher who claim that we can and must find a different way of expressing this sentiment, "creat[ing] a new social world with

gentler norms" are both being "sentimental and utopian" and also thereby engaging in a kind of philosophical over-reach. It is a matter of observable convention that we *do* express these sentiments by means of punishing. Whether it is possible that we might come to express them by other means is not for the philosophers to say. "In thinking normatively about the social world," Kleinfeld tells us, "understanding should come first, advocacy second, and wishful thinking should have no place at all."[148] [149]

Finally, Kleinfeld attempts to undercut the general moral picture that he takes to provide the positive basis for the skeptical position: namely, that suffering is bad, and ought to be avoided whenever possible. "[The] fixation on justifying punishment," Kleinfeld writes

> "…lends criminal theory a certain Enlightenment-humanist emotional flavor, for its spiritual root is revulsion at suffering…There are other ways of thinking of moral life than this; the moral and political thinkers of the ancient world, for example, would never have thought *pain* to be of such overwhelming moral importance."[150]

We should not, he thinks, treat the infliction of suffering as a *prima facie* wrong, requiring a special justification as such, any more than we should treat the giving of pleasure as a *prima facie* good. Punishment "is not," then, "a *prima facie* evil but, where used appropriately to secure the right, simply the language in which one expresses a commitment to what is moral."[151]

---

[148] *Ibid*. p1487.

[149] This view of the philosopher's proper place in a discussion about the moral basis of punishment is grounded in a more general view about the proper place of the philosopher. According to this view, an important object of normative philosophy is to bring the immanent values that make up this embodied ethical life to light, to render them explicit. That is, the role of philosophy is not chiefly to define and defend some set of abstract or a priori ideals, which are then applied to the world to dictate how it should be ordered, but to rationally reconstruct the normative order already at work in the world in order to see that normative order more clearly and critique it. The philosopher thus stands in an interpretive rather than a concept-application relationship to the social world" (Kleinfeld p1487).

[150] *Ibid.* p1498.

[151] *Ibid.* p1514.

Kleinfeld questions, in short, what seemed to be the one proposition on which all philosophers of punishment were in agreement: That punishment is by its nature the kind of thing that requires a special justification. The fact that punishment involves suffering, Kleinfeld thinks, is not itself a reason to think that there is any big justificatory problem about punishment.

For any of these three responses, there is no doubt some strain of skepticism about punishment it stands to rebuke, and probably for anyone inclined to the skeptical worry, these responses offer some form of helpful corrective. Even taken together, though, these replies will not answer the challenge posed by Kleinfeld's best skeptical critic. This critic is not concerned with suffering, but with rights—particularly rights of autonomy. The deep problem about punishment is not, on her view, that it causes of suffering, but that it seems to constitute the violation of that to which autonomous moral agents have a right—in the case of state punishment often the basic rights of freedom of movement, association, and even the right not to be killed. This critic is not naively utopian or sentimental in her assessment of what is possible, but notes there has *in fact* been social progress that moves us to express by other means what we once expressed by punishing, and where we do still employ punishment to this end, to use less harsh and brutish forms. (Recall Flew's claim that where punishment once meant the stockade and gallows, it may now mean community service.) She will note, too, that philosophers as "advocates" seem in fact to have played a role in larger social movements both toward and away from the employing of harsh or retributive forms of punishment as a means of social control and expression.[152] Finally, this critic will not be satisfied with the claim that punishment "just is", as a matter of convention, how we express our commitment to violated moral norms and solidarity with victims. If she is to surrender her commitment to the "sentimental and utopian" idea that

[152] Freed 1992, Stith 1998, Whitman 2003.

another world is possible (and that she, as a philosopher, might have some effective and legitimate role to play in bringing that about), and she is to do so despite what looks to be good evidence that it might well be, Kleinfeld had better be prepared to offer her not just a sociological or linguistic account of how punishment *did in fact* come to be the way we express our commitment and solidarity, but a philosophical account of why our commitment and solidarity *must* be expressed in this way—or of why, at least, punishment is *by its nature* the *best* way of expressing it.

Kleinfeld tells us that punishment just is our conventional means of expressing a commitment to the abstract norms wrongdoing violates, including the norms of dignity and equality that we violate in treating other's rights as violable and their interests as less important than our own. This just is how human beings "say" these things, and the fact that this way of saying it typically involves suffering gives us no special reason for demanding a special justification. But there is more to say. That we express this set of ideas by punishment is not a mere accident of history. Neither is it merely the bare fact of our aversion to loss or suffering leveraged for the purpose of communicating our aversion to wrongdoing. There is a deeper, more complex story here between punishment's form and its expressive content—one that merits greater thought and explication.

If this is so, it might give us some reason to conclude that while our personal, social, and political practices of punishing may change over time in terms of the forms of deprivations and burdens we tend to impose and the frequency with which we impose them, that we will always (or at any rate *should* always) engage in some forms of punishing. But even if the critic remains unconvinced that punishment is an indispensible tool for repairing the damage wrongdoing sometimes does to our shared moral life, might still be brought around if we could convince her

that punishment does not necessarily violate any right of the punished. Such an argument may be necessary to convince even the critic who feels there *is* something to the idea that punishment in some form or another is indispensible. To say that there are vital human needs that punishment may in some cases work uniquely to serve is not necessarily to say that we are entitled to punish. There are some things I am not morally entitled to do even to save my own life, however valuable my life may be.

So why think that punishing someone is consistent with the proper regard we owe them? Reconstructivism has an advantage over deterrence theories because it does not treat the interests of the wrongdoer as something we simply tradeoff for the sake of others' interests. What we compromise them for the sake of, on this theory, is a form of group solidarity built around a shared set of moral norms, embodied in a shared moral culture—one that the members, including the wrongdoer, not only benefit from, but (according to Kleinfeld) they require for their personal well-being.[153] This view may sound familiar. In chapter 3 I argue for a similar account of punishment's value in the context of our personal relationships. In that case the context was not one of a complex society, but of a dyadic relationship.

If punishment does not instrumentalize the punished in either of the two important senses, is there anything left for the critic to worry about? Yes. She may still be concerned to know if punishment involves the violation of any right that the punished might have. This concern may have seemed less serious when it came up in the context of our interpersonal relationships, where the punished is deprived by her friend of what she only has, to begin with, in virtue of being in this relationship, which both parties are more or less free to leave. As a friend I may owe you a certain amount of my time and attention, but I do not owe you my friendship. In the case of state

[153] Kleinfeld 2016, p1493.

punishment, by contrast, the kinds of deprivations the state is in a position to impose are typically of a different nature, sometimes involving a forced limitation upon our freedom of movement or seizure of property.

There is an argument to be made that the critic is wrong in thinking that we can live out our commitment to the right kind of shared moral norms without punishing. Kleinfeld, though, has located the problem in the wrong place. The problem is not that the critic is being naïve in her assessment of our human capacity to live up to an ideal, nor that she is engaging in a kind of disciplinary over-reach. The problem is that this is wrong, even in principle. We cannot live out our moral commitments without punishment, and the mistake of thinking we could is one that it is well within the job description of the philosopher to diagnose. To aspire to "a world altogether without hard treatment" (if, indeed, anyone really does aspire to this) is to make at least one of the following three mistakes: To take too narrow a view of what counts as "hard treatment"; to hold that there is no necessary connection between punishment's form and expressive content; to adopt the wrong ideal of human life. If, on the other hand, we understand punishment in the broad terms I have suggest, understand even an ideal of human life as involving moral failure and subsequent conflict, and if we understand that conflict, when it takes the form of punishment, to be an indispensible tool for repairing what wrongdoing breaks, then it makes more sense to conclude that punishment in some form or other will always be with us.

One of the premises from which this dissertation in general and the arguments of each individual chapter have begun is that people— even idealized and operating under idealized conditions, both personal and political— will sometimes fail to do what they should, and that wrong-reactive attitudes and emotions like anger, resentment, disappointment, and frustration will sometimes be the appropriate response to such failure. Some of these instances of moral

failure will not be of the sort that we can appropriately overlook or respond to with warmth and good-humor. Interpersonal conflict is therefore a natural and characteristic feature of human life. So conflict is a matter of course in even the best lives, lived by the kindest, bravest, and most attentive and fortunate among us. This is not to say that people are by nature cruel or selfish, but only that we are the kinds of things that get tired and hungry, that learn boundaries by trespassing them, that struggle and often fail to understand. So often we fail to understand what we should do, fail to do what we know we should, fail to attend to the relevant facts about the people we love, the circumstances we are in, and what they call for. These are not the inevitable results of a bad or fallen nature, but just of being the kinds of things with finite resources of time and attention. That we are things with finite resources of time and attention is not the kind of thing we should abstract away from in our idealizing. A proper ideal of human life is one in which we perhaps fail less and for different reasons. But even in such a world, resentment, anger, and disappointment will sometimes be appropriate, as will their expression in the context of the sorts of (often enormously) difficult exchange by which we work together through what has happened.

One might agree that the ideal to which we should strive, as a society and in our personal relationships, will involve both moral error and treatment responsive to moral error that constitutes a deprivation or burden, just not the kind of burden or deprivation one ought to call punishment, *per se.* Such a person might equate punishment with, e.g., incarceration in the case of state punishment, spanking in the case of childhood punishment, or overly-harsh and drawn-out forms of emotional withholding in the context of a friendship. In chapter's 1, 2, and 3, respectively, I argued against understanding punishment as limited to these forms of treatment. Jean Hampton, one of the handful of philosopher's Kleinfeld places at center of the reconstructivist tradition, suggests a similarly broad account of what might count as

punishment.[154] We have reason to adopt this more capacious view of what punishment *is* that do not presume the correctness of any particular normative theory of that (set of) practice(s).

If one accepts the right account of what punishment is (a broad one), and of the sort of ideal of human life to which we should aspire (one that includes the experience and expression of the wrong-reactive attitudes), one cannot, I think, conclude that punishment is the sort of thing past which human society's either could evolve, or should aspire to. One might well, of course, conclude that incarceration, spanking, pettiness and emotional abuse are the sorts of things past that we might, as a society, evolve. One might well think we ought to. But one does not thereby hold the view that we ever will, or *could*, or *ought to*, "evolve past" punishing more broadly.

Kleinfeld's view is a crucial move in the right direction. In identifying solidarity as the "lodestar normative concept" in a proper theory of punishment, he implicitly puts relationships at that theory's heart. Solidarity is something we are said to *show to* or *stand in with* others, a unifying way of being related to them. It is a way of being unified with others that can be powerful, even without being intimate. Solidarity is something I can stand in with those whom I have never met, and those for whom I have a strong personal distaste. I don't have to know or like someone to stand in solidarity with her. While never mentioning relationships explicitly in his initial account of reconstructivism, he brings them front and center, replacing desert of the individual or happiness of the aggregate with solidarity among persons as the final, justifying aim of state punishment.

---

[154] Hampton, in particular, proposes the phrase "disruption of the freedom to purse the satisfaction of one's desires" as the proper characterization of punishment, which she prefers because it (A) does not imply that punishment necessarily involves any pain, per se, (just, typically, the discomfort autonomous agents naturally feel at having their autonomy impeded) and (B) that though it involves some loss of freedom it can be the loss of freedom involved not just in e.g. incarceration, but in, e.g., having one's allowanced stopped, or even being made to endure a lecture (Hampton 1984, p224).

In this and several other key ways, Kleinfeld's reconstructivism invites a relationship-centered approach to the question of state punishment, though it fails to offer one. It is an account of our more general political and social relations with one another in which solidarity ultimately figures, and can provide us with a unified way of understanding *both* what should constitute criminal wrongdoing, and what constitutes an appropriate response to it. (One of Kleinfeld's central complaints about the existing literature is that it that too little of it is dedicated to understanding what it *is* to do wrong, as a pre-requisite for understanding what an adequate response to wrongdoing will look like, and what value such a response might serve).

But in insisting on what is at bottom a descriptive, naturalistic picture of our commitment to "embodied ethical life", Kleinfeld leaves us with insufficient tools for offering a robust, convincing account of punishment's place and value. It is not enough to say, at bottom, "this is just how we do it, and how we have to do it if we want things to go as we would prefer that they do." The claims that solidarity has and makes on us are not merely contingent on a further, personal commitment to well-being (whether individual or collective). Solidarity doesn't just make moral claims on us, but *has* a moral claim on us. It is something we can owe to one another, even when we wish it were otherwise. We may (and do, I believe) also have reasons for expressing solidarity in the ways that we do which are not merely conventional. If we begin from the right *normative* account of the relationship, of the claims it has and makes on us, and of the ways in which we are called both to lay and to answer those claims, we will have provided ourselves with a better set of resources for answering Kleinfeld's best critic in a way that might satisfy her, finally putting the skeptical worry to rest.

A better defense of criminal punishment will begin from an account not just of what the criminal law and criminal wrongdoing are, but with the *prior* question of what an ideal of the

relationship between citizen and state would look like. From here we would have the resources to build not only an account of what wrongdoing in this context really is (a question that Kleinfeld, too, is concerned with), but of what responses to wrongdoing are consistent with the rights and obligations that structure the relationship in which the wrong occurred. An account of the relationship would also give us a way of beginning to answer the question of what *repair* in the context of such a relationship *would be*, and, subsequently, to begin to assess the question of whether or not and how any particular convention we might have engaging in such repair might turn out to be effective.

E. Conclusion

When we start from the right kind of relationship we are then in a position to develop the right kind of relationship-centered account of criminal punishment. This will be an account consistent with the idea of criminal punishment and the criminal law more generally as a distinctively *public* enterprise, whose value is as a public manifestation of an institution-wide commitment to a basic scheme of political values and valuing. It will be an account on which the public commitment it manifests is one of solidarity. Solidarity is not, though, a way of being situated relative to a set of norms, "embodied" or otherwise, but a way of relating to other *people*. It is, additionally, not the sort of attitude that we must find ways of taking up only toward victims, or toward other participants in good standing, but toward offenders, as well. The distinctive contribution that the relationship-centered approach might make emerges when we start by centering the relationship between offender and offended, where in this case the offense we are worried about is a kind of offense against the state. Part of punishment's distinctive value (where it has any) will lie in its potential to repair that relationship, and thus serve to further the

value that this (institutional) relationship has for us. It is important, then, that neither the *state*, on the one hand, nor the *offender*, on the other, should fall out of our account, or turn out to play only some indirect or purely instrumental role there.

Much more is required of a relationship-centered account of state punishment , of course, than that it start from an account of the right relationship. Such an account would need to address the questions that I have said that any account must: namely, the questions of *boundary, aim, method, authority*, and *permissibility.* I will conclude with some initial speculations about how the right kind of relationship-centered account of criminal punishment might handle these questions.

Boundary. The relationship-centered account may not provide an absolutely unique route into saying anything we would especially like to say about what does and does not constitute criminal punishment. It does, I have suggested, do a better job than, e.g., restorative justice models at helping to carve out the boundary between criminal punishment and other forms of punishment in which the state may be involved (most notably civil sanctions). The sort of relationship-centered account of punishment I have suggested provides a way for us to think about the distinctively *public* aspect of criminal punishment and the network of normative relationships it takes place in the context of, while still providing us with the resources for doing at least one of the things that it seems to me restorative justice models do really well: Broadening our sense of the forms of depriving and burdening that might be imposed as punishment. In the proceeding chapters I have stressed the point that we should not limit our sense of which forms of hard treatment do and do not count as punishments to those forms of hard-treatment typically imposed by the state. The restorative justice theorist reminds us that we should not allow that experience to limit our sense of which forms of hard treatment might count as punishment by the

state either. Just because an imposition is in some sense gentler that those to which we are accustomed, does not mean it doesn't amount to a form of punishment. The relationship-centered account of punishment I have sketched has the resources to take on and advance this crucial insight of the restorative justice model, without having to take on that model's limitations.

Aim. I have said as a general matter that the aim of punishment will be to advance the aims of the relationship inside of which it occurs, contributing to the repair of the damages done by the wrongdoing to which it constitutes a response. The aim of criminal punishment in particular, I have claimed, is the repair of the relationship between criminal offenders and the state, providing the offender with a viable path back to good standing.

This is not to say that punishment will serve no crucial function other than the repair of this particular relationship. There is surely something right in the thought that prosecution and punishment of criminal offenders plays an important role in ensuring the "social dignity" of crime victims. There is certainly a case to be made that the failure to prosecute and punish some violent crimes, at least—and *especially* to prosecute and punish such crimes when committed against citizens of one sort but not another—is to treat the lives of victims as cheap. To claim that punishment's distinctive justificatory aim is to repair the relationship between the offender and the state is not to claim, either, that punishment alone is always sufficient or even always necessary to achieve that aim. There are cases in which it seems plausible to think that an offender's breach is one that can be repaired by paying a fine, and even one's in which a display of sincere remorse before a judge might be sufficient. My claim is just that where punishment *is* defensible, it will, among other things, aim to repair the state-offender relationship.

Method. In order for such an account to be theoretically useful, we would need to spell out a set of methods and mechanisms whereby punishment works to serve the relevant reparative

aim. It would need to tell us something, in other words, about *how* punishment could work to repair the relevant kind of relationship, potentially bringing the offender back into a place of good standing in terms of his relationship to the state. Will punishment here function communicatively, as it does in the case of friendship, by "telling" the offender something that he must know in order for repair to be possible? Perhaps, but in the case of friendship, which is a relatively intimate relationship, by its nature involving certain forms of shared communicative capacities, we have better reasons for thinking so. Does citizenship—properly idealized at any rate—involve such some shared communicative capacity amongst citizens? Perhaps punishment here works by some entirely different mechanism, one more appropriate to an impersonal sort of relationship. A full relationship-centered account of punishment would have to say, and the answer it offers would need to be such that it does not run us into the same old problem that always arises for instrumental theories of state punishment in this tradition: that, in practice at least, the methods specified operate (where they are effective at all) by bypassing, or even operating *on* rational capacities rather than engaging them.

Authority. The shape and limits of the state's authority to punish have been central issues in the literature. What the relationship-centered account adds is to have us notice that when we talk about the state's authority, we are always talking about something that only exists as a feature of a relationship or set of relationships it has to others. Kleinfeld claims that too much ink has already been spilled on the question of what the nature and limits of the state's authority to punish might be—particularly where this work has been to the exclusion of work on what seems to him to be the more fundamental question: *What is wrongdoing? What kind of damage does it do?* "[B]y ignoring or minimizing the question of what crime and wrongdoing are," he says, criminal theory "deprived itself of the resources by which to answer its own question"—the

question of what right the state has to punish. "The wrongdoing-redress structure," he continues, "is not a matter for a discrete and isolated subfield; it is a basic feature of human social organization. It is not even in the first place a matter of law or the state."[155]

The thought that we might learn something new and relevant about the wrong-redress dyad by investigating it across a range of human contexts is the motivating thought behind this entire dissertation, and I began this work with much sympathy for Kleinfeld's position. But having taken up this question of what the broader place of wrongdoing and redress in human life might be, it has turned out that the first step in answering *either* question must involve a *general specification of the kind of relationship that wrongdoing and redress are taking place in*. What constitutes wrongdoing changes from one kind of relationship to another, as does the kind of damage that wrongdoing involves, as does the kind of response to wrongdoing that will be defensible. It could not be otherwise. All of this follows naturally from the fact that each kind of human relationship has distinctive aims and value, and involves a distinctive set of rights and obligations. To investigate the nature and limits of the authority that the state has in our lives, then, does in fact turn out to be the first question that requires settling—a question conceptually prior to the question of what wrongdoing amounts to in that context.

The relationship-centered account of punishment demands, along with Kleinfeld, that we understand philosophy of (state) punishment as inextricable from political philosophy more generally, but invites us to see political philosophy itself as a way of thinking about a set of relationships—not relationships between private persons, but between persons and the state, or between persons qua citizens of that state. The relationship-centered account situates the question of authority in the context of a broader set of questions about the norms governing a

[155] Kleinfeld 2016, p1503.

particular kind of relationship—a set of questions that, once settled, will help us to settle in turn both what the nature and limits of the authority of any given member of that relationship might look like, and what it is to do wrong in the context of that relationship.

Permissibility. There is much work here still left to do—too much for me to be able to draw a set of concrete conclusions about what would and would not count as permissible punishment on a relationship-centered view of punishment. There are, though, a few things I think we can say with some certainty. The relationship-centered view of punishment is fundamentally a reparative view, on which authority to punish is always grounded in a commitment to the relationship itself, and on which punishment's justification always rests on its usefulness in advancing those interests in the wake of the damage done by wrongdoing. So what will not be allowed on a relationship-centered theory of punishment are those forms of punishment that constitute a permanent impairment of the relevant relationship. In the context of state punishment, it looks, then, like, e.g., the death penalty is highly unlikely to meet that bar. When the state kills its citizens, it is very hard to see how this could amount to a way of repairing the relationship between citizen and state. As a punishment, execution not only fails to constitute a form of repair, but to rule out the possibility of repair all together. Likewise, we might consider cases such as the loss of voting rights by felons or a sentence of life without parole.[156] Insofar as permissible punishment will aim to preserve and repair the relationship that wrongdoing damages, punishments that impair and degrade the relationship, especially in its most

[156] It may be less obvious that life without parole (LWOP) constitutes the kind of objectionable rupture in the relationship between citizen and state that execution or disenfranchisement does. I believe that it does, but it requires further argument to show why. For an excellent discussion of the topic, see Sharon Dolovich's "Creating the Permanent Prisoner," in which she argues that the use of LWOP contributes to "the pervasive normative conception of prisoners as both noncitizens and nonhumans" (Dolovich 2012, p96).

fundamental aspects (e.g. voting rights in the context of democratic citizenship), will miss the mark.

On the right kind of relationship-centered account of punishment, it will always be the burden of those who aim to justify any particular form of punishment to show that it meaningfully advances the aim of repair and reintegration. In starting from a picture of what an ideal of the relevant form of relationship would look like, it provides us with the resources we need to investigate all the important questions that follow: What counts as a violation of that relationship? How does (some particular instance of) wrongdoing damage the relationship? What sorts of subsequent interventions might effectively work to set things right again? The theory of punishment that should satisfy us will be one concerned to investigate these questions, directing our attention to them, and offering us something new to say in response. This is the work I hope the right kind of relationship-theory of punishment might do.

Bibliography

Alexander, Jennifer & Valdovinos, Maria. 2011. "Punishment." *Encyclopedia of Child Behavior and Development*. Sam Goldstein & Jack Naglieri, eds. 1202.

Almaatouq, Abdulla; Radaelli, Laura; Pentland, Alex; Shmueli, Erez. 2016. "Are You Your Friends' Friend? Poor Perception of Friendship Ties Limits the Ability to Promote Behavioral Change." *PLoS ONE*. 11:e0151588.

Aristotle. 2009. *The Nichomacean Ethics.* Lesley Brown, ed. W.D. Ross, trans. Oxford University Press.

Armstrong, K.G. 1961. "The Retributivist Hits Back." *Mind*. 70:471-490.

Aufhauser, Marcia Cavell. 1975. "Guilt and Guilt Feelings: Power and the Limits of Power." *Ethics.* 85:288-297

Benn, S.I. 1958. "An Approach to Problems of Punishment." *Philosophy*. 33:325-341.

Bennett, Christopher. 2008. *The Apology Ritual: A Philosophical Theory of Punishment*. Cambridge University Press.

Berman, Mitchell. 2008. "Punishment and Justification." *Ethics*. 118:258-290.

Boonin, David. 2008. *The Problem of Punishment*. Cambridge University Press.

Bradshaw, John. *Healing the Shame That Binds You.* 2005. Health Communications, Inc.

Braithwaite, John. 1989. *Crime, Shame and Reintegration*. Cambridge University Press.

—. 2001. *Restorative Justice and Responsive Regulation*. Oxford University Press.

Braithwaite, John & Pettit, Philip. 1990. *Not Just Deserts*. Oxford University Press.

Brooks, Thom. 2007. "Review: Punishment and Retribution by Leo Zaibert." *New Criminal Law Review: An International and Interdisciplinary Journal*. 10: 311-314.

Broucek, F.J. 1982. "Shame and Its Relationship to Early Narcissistic Development." *The International Journal of Psycho-Analysis.* 63:369–378.

CERA (Communities Embracing Restorative Action). 2017. cerasociety.org.

Clark, Andy. 2000. "Word and Action: Reconciling Rules and Know-How in Moral Cognition." *Canadian Journal of Philosophy*. 26:267-289.

Cocking, Dean & Kennett, Jeannette. 1998. "Friendship and the Self." *Ethics*. 108, 504-527.

Cragg, Wesley. 1992. *The Practice of Punishment: Toward a Theory of Restorative Justice*. Routledge Press.

Delaney, Rob. 2013. *Rob Delaney: Mother. Wife. Sister. Human. Warrior. Falcon. Yardstick. Turban. Cabbage.* Spiegel & Grau.

Dillon, Robin S. 2001. "Self-Forgiveness and Self Respect." *Ethics*. 112:53-83.

Dolinko, David. 1991. "Some Thoughts About Retributivism." *Ethics*. 101: 537-559.

—. 1993. "Three Mistakes of Retributivism." *UCLA Law Review*. 39:1623-1657.

Dolovich, Sharon. 1999. "Cruelty, Prison Conditions and the Eighth Amendment." *New York University Law Review*. 84: 881-979.

—. 2004. "Legitimate Punishment in Liberal Democracy." *Buffalo Criminal Law Review*. 7:307-442.

—. 2011. "Exclusion and Control in the Carceral State." *Berkeley Journal of Criminal Law*. 16:259-339.

—. 2012. "Creating the Permanent Prisoner." *Life Without Parole: America's New Death Penalty?* Charles Ogletree and Austin Sarat, eds. New York University Press.

—. 2012. "Two Models of the Prison: Accidental Humanity and Hypermasculinity in the L.A. County Jail." *Journal of Criminal Law and Criminology*. 102:965-1118.

Donelson, Raff. 2016. "Cruel and Unusual What? Toward a Unified Definition of Punishment." *Washington University Jurisprudence Review*. 9:1-41.

Duff, R.A. 1986. *Trials and Punishments*. Cambridge University Press.

—. 1988. "Punishment and Penance." *Aristotelian Society: Supplementary Volume.* 62:153-167.

—. 2001. *Punishment, Communication, and Community*. Oxford University Press.

Dwyer, Susan. 1999. "Moral Competence." *Philosophy and Linguistics*. Kumiko Murasugi and Robery Stainton, eds. 169-190.

—. 2003. "Moral Development and Moral Responsibility." *The Monist*. 86:181-199.

Feinberg, Joel. 1965. "The Expressive Function of Punishment." *The Monist.* 49:397-423.

—. 1984-1988. *The Limits of Criminal Law*. Oxford University Press.

Flew, Anthony. 1954. “The Justification of Punishment.” *Philosophy*. 29:291-307.

Freed, Daniel. 1992. “Federal Sentencing in the Wake of Guidelines: Unacceptable Limits on the Discretion of Sentencers.” *Yale Law Journal*. 101: 1681-175.

Gershman, Bennett. 2005. “Prosecutorial Ethics and Victims’ Rights: The Prosecutor’s Duty of Neutrality.” *Lewis and Clark Law Review*. 9:559-579.

Gill, Frances E. 2003. *The Moral Benefit of Punishment: Self-Determination as a Goal of Correctional Counseling*. Lexington Books.

Glick, Ephraim. 2012. “Abilities and Know-How Attributions.” In *New Essays on Knowledge Ascriptions*. Jessica Brown & Mikkel Gerken, eds. Oxford University Press.

Goldstein, Abraham. 1982 .“Defining the Role of the Victim in Criminal Prosecution.” 52:515-561.

Halford, Macy. 2009. “Read This Book If: You Want Your Child To Be Perfect.” *The New Yorker.* November 12, 2009.

Hampton, Jean. 1984. “The Moral Education Theory of Punishment.” *Philosophy & Public Affairs.* 13:208-238.

—. 1992. “An Expressive Theory of Retribution.” *Retributivism and Its Critics.* Ed. Wesley Cragg. Franz Steiner Verlag.

—. 1994. “The Common Faith of Liberalism.” *Pacific Philosophical Quarterly*. 75:186-216.

Harrison, Ross. 1988. “Punishment and Crime.” *Aristotelian Society Supplementary Volume*. 62:139-167.

Hart, H.L.A. 1960. “Prolegomenon to the Principles of Punishment.” *Proceedings of the Aristotelian Society*. 60:1-26.

—. 1968. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press.

Hegel, G.W.F. 1991. *Elements of the Philosophy of Right*. Allen Wood, ed. H.B. Nisbet, trans. Cambridge University Press.

Helm, Bennett. 2008. “Plural Agents.” *Nous*. 42:17-49.

Herman, Barbara. *Forthcoming.* “Religion and the Highest Good: Speaking to the Heart of Even the Best of Us.” In *Freedom and Spontaneity in Kant.* Kate Moran, ed. Cambridge University Press.

—. 1993. *The Practice of Moral Judgment.* Harvard University Press.

Hieronymi, Pamela. 2001. "Articulating an Uncompromising Forgiveness." *Philosophy and Phenomenological Research*. 62:529-555.

Hochschild, Adam. 2005. *Bury the Chains: Prophets and Rebels in the Fight to Free an Empire's Slaves.* Macmillan.

Holmgren, Margaret R. 1999. "Self-Forgiveness and Responsible Moral Agency." *Journal of Value Inquiry*. 32:75-91.

Johnstone, Gerry & van Ness, Daniel, eds. 2007. *Handbook of Restorative Justice*. Taylor and Francis.

Kaeble, Danielle & Glaze, Lauren E. 2015. *Correctional Populations in the United States*. US Bureau of the Justice Statistics.

Kant, Immanuel.1899. *Kant on Education*. Annette Churton, trans. Kegan Paul, Trench, Trubner & Co Ltd.

—. 1999. *Critique of Pure Reason*. Paul Guyer & Allen Wood , eds. & trans. Cambridge University Press.

—. 1999. *Religion Within the Boundaries of Mere Reason*. Allen Wood & George Di Giovanni, eds. & trans. Cambridge University Press.

—. 2012. *Groundwork of the Metaphysics of Morals*, *2nd ed*. Mary Gregor & Jen Timmermann, eds & trans. Cambridge University Press.

Karlsson, Gunnar & Sjoberg, Lennart Gustav. 2009. "The Experiences of Guilt and Shame: A Phenomenological-Psychological Study." *Human Studies*. 32:335-355.

Karson, Michael. 2014. "Punishment Doesn't Work." *Psychology Today.* January 2014.

Kelly, Erin. 2009. "Criminal Justice Without Retribution." *Journal of Philosophy*. 106: 440–462.

Khazan, Olga. 2016. "No Spanking, No Time-Out, No Problems." *The Atlantic*. March 28, 2016.

Kleinfeld, Joshua. 2016. "Reconstructivism: The Place of Criminal Law in Ethical Life." *Harvard Law Review*. 129:1485-1565.

Lacey, Nicola. 1988. *State Punishment: Political Principles and Community Values*. Routledge.

Lewis, Katherine Reynolds. 2015. “What If Everything You Knew About Disciplining Kids Was Wrong?” *Mother Jones*. July/August 2015.

Locke, John. 1779. *Some Thoughts Concerning Education.* J. & R. Tonson.

Mabbott, J.D. 1955. “Professor Flew on Punishment.” *Philosophy*. 30:256-265.

Mangasarian, M.M. 1894. “The Punishment of Children.” *International Journal of Ethics*. 4:493-498.

Singer, Marcus. 1959. “On Duties to Oneself. ” *Ethics*. 69:202-205

Margalit, Avishai. 1998. *The Decent Society*. Naomi Goldblum, trans. Harvard University Press.

Marshall, J.D. 1972. “On Why We Don’t Punish Children.” *Journal of Educational Philosophy and Theory*. 4:57-68.

—. 2007. “Punishment and Education.” *Educational Theory*. 4:57-68.

Matychuk, Paul. 2004. “The Role of Child-directed Speech in Language Acquisition.” *Language Sciences.* 27: 301-379.

McCloskey, H.J. 1962. “The Complexity of the Concepts of Punishment.” *Philosophy*. 37:307-325.

McPherson, Thomas. 1967. “Punishment: Definition and Justification.” *Analysis*. 28:21-27.

Morris, Allison & Maxwell, Gabrielle, eds. 2001. *Restorative Justice for Juveniles: Conferencing, Mediation and Circles*. Hart Publishing.

Morris, Herbert. 1968. “Persons and Punishment.” *The Monist*. 52:475-501.

—. 1971. “Guilt and Suffering.” *Philosophy East and West*. 21:419-434.

—. 1981. “A Paternalistic Theory of Punishment.” *American Philosophical Quarterly*. 18:263-271.

Morrison, Andrew. 1983. “Shame, Ideal Self, and Narcissism.” *Contemporary Psychoanalysis*. 19:295-318.

—. 1997. *Shame: The Underside of Narcissism.* Routledge Press.

Muchnik, Pablo. 2014. “The Heart as Locus of Moral Struggle in the *Religion.*” *Kant on Emotion and Value*. Alix Cohen, ed. Palgrave Macmillan. 224-244.

Murphy, Jeffrie. 1973. “Marxism and Retribution.” *Philosophy and Public Affairs*. 2:217-243.

—. 1988. “Jean Hampton on Immorality, Self-Hatred, and Self-Forgiveness.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 89: 215-236.

—. 1999. “Shame Creeps Through Guilt and Feels Like Retribution.” *Law and Philosophy*. 18:327-344.

—. 2007. “Legal Moralism and Retribution Revisited.” *Criminal Law and Philosophy*. 1:5–20.

Nathanson, Donald. 1987. *The Many Faces of Shame*. Guilford Press.

Newton, John. 1793. *Letters to a Wife, by the Author of Cardiphonia*. J. Johnson, Publisher.

—. 1868. *John Newton, of Olney and St. Mary Woolnoth: An Autobiography and Narrative, Compiled Chiefly from His Diary and Other Unpublished Documents, 2nd edition*. Josiah Bull, ed. The Religious Tract Society.

—. 2003. *Out of the Depths.* Dennis Hillman, ed. Grand Rapids Press.

Nozick, Robert. 1981. *Philosophical Explanations*. Harvard University Press.

Nussbaum, Martha. 2006. *Hiding from Humanity*: *Disgust, Shame, and the Law*. Princeton University Press.

Atonement. 1971 *The Compact Edition of the Oxford English Dictionary.* Oxford University Press. 539.

Primoratz, Igor. 1989. “Punishment as Language.” *Journal of Philosophy*. 64:187-205.

Quinn, Warren. 1985. “The Right to Threaten and the Right to Punish.” *Philosophy and Public Affairs*. 14: 327-373.

Quinton, A.M. 1959. “On Punishment.” *Analysis*. 20:10-13.

Radzik, Linda. 2004. “Making Amends.” *American Philosophical Quarterly*. 41:141-154.

Rawls, John. 1955. “Two Concepts of Rules.” *The Philosophical Review.* 64:3-32.

—. 2000. *Lectures on the History of Moral Philosophy*. Barbara Herman, ed. Harvard University Press.

Rorty, Amelie Oksenberg. 1986. “The Historicity of Psychological Attitudes: Love Is Not Love Which Alters Not When It Alteration Finds.” *Midwest Studies in Philosophy*. 399-412.

Ryle, Gilbert. 1945. “Knowing How and Knowing That.” *Proceedings of the Aristotelian Society*. 56:1-16.

Scanlon, T.M. 2008. *Moral Dimensions*. Harvard University Press.

—. 2000. *What We Owe to Each Other*. Belknap Press.

Scheid, Don E. 1980. “Note on Defining ‘Punishment.’” *Canadian Journal of Philosophy*. 10:456.

—. 1997. “Constructing a Theory of Punishment, Desert, and the Distribution of Punishments.” *Canadian Journal of Law and Jurisprudence.* 10:441–506.

Shafer-Landau, Russ. 1996. “The Failure of Retributivism.” *Philosophical Studies*. 82:289-316.

Shapiro, Tamar. 1999. “What is a Child?” *Ethics*. 109:715-738.

Sherman, Nancy. 1987. “Aristotle on Friendship and the Shared Life.” *Philosophy & Phenomenological Research.* 47:589-613.

Shiffrin, Seana Valentine. 2014. *Speech Matters: On Lying, Morality, and the Law.* Princeton University Press.

Skillen, A.J. 1980. “How to Say Things with Walls.” *Philosophy.* 55:509-523.

Smetana, Judith. 1999. “The Role of Parents in Moral Development: A Social Domain Analysis.” *Journal of Moral Education*. 28:311-321.

Stith, Kate & Cabranes, Jose. 1998. *Fear of Judging: Sentencing Guidelines in the Federal Courts*. University of Chicago Press.

Strawson, P.F. 1962. “Freedom and Resentment.” *Proceedings of the British Academy.* 48:1-25.

Sussman, David. 2005. “Perversity of the Heart.” *Philosophical Review*. 114:153-177.

Telfer, Elizabeth. 1970. “Friendship.” *Proceedings of the Aristotelian Society*. 71:223-241.

Thomas, Laurence. 1987. “Friendship.” *Synthese*. 72:217-236.

—. 1989. “Friends and Lovers.” *Person to Person*. George Graham & Hugh La Follette, eds. Temple University Press.

—. 1993. “Friendship and Other Loves.” *Friendship: A Philosophical Reader*. Neera Kapur Badhwar, ed. Cornell University Press.

von Hirsch, Andrew. 1993. *Censure and Sanction*. Oxford University Press.

Walker, Lawrence J. 2002. "Moral Exemplarity." *Bringing in a New Era in Character Education*. William Damon, ed. Stanford University: Hoover Institution Press. 65-83.

Weber, Max. 1978. *Economy and Society: An Outline of Interpretive Sociology*. Guenther Roth & Claus Wittich, eds. & trans. University of California Press.

White, Richard. 2001. *Love's Philosophy*. Rowman & Littlefield.

Whitman, J.Q. 2003. "A Plea Against Retributivism." *Buffalo Criminal Law Review*. 7:85-107.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy.* Harvard University Press.

Wolf, Susan. 2011. "Blame, Italian Style." R. Jay Wallace, Rahul Kumar, & Samuel Freeman, eds. *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*. Oxford University Press. 332-247.

Wood, Allen. 1970. *Kant's Moral Religion.* Cornell University Press.

—. 1991. "Kant's Deism." *Kant's Philosophy of Religion Reconsidered.* Philip J Rossie & Michael J. Wreen, eds. 1-21.

Wootton, Barbara. 1963. *Crime and the Criminal Law*. Stevens Press.

Yao, Vida. 2016. "Grace: Goodness in Loving the Bad." In *Loving the Bad & Not Giving a Damn: A Defense of Psychic Disharmony.* Dissertation. University of North Carolina, Chapel Hill.

Zaibert, Leo. 2006. *Punishment and Retribution*. University of California Press.

Zupancic, Melissa & Kreidler, Maryhelen. 1999. "Shame and the Fear of Feeling." *Perspectives in Psychiatric Care*. 35:29-34.