Forthcoming in S. Figueroa Rubio and I. Ortiz de Urbina (Eds.). *Opresión, responsabilidad y delito*. [*Oppression, Responsibility, and Crime*] Madrid, Spain: Marcial Pons.

Blame, Oppression, and Retributive Punishment

Manuel Vargas
University of California San Diego

There are different ways to err in blaming. First, blaming might go wrong individually, as a matter of an individual, token-level judgment. I might wrongly blame you, or fail to blame you, because of a mistaken belief about you or your action. Call this a *token error*. Second, an instance of blame might be in error if blame itself is essentially in error. This might happen if deserved blame always requires something impossible or unavailable to creatures like us. If blame is only justified when self-creating agents act wrongly with the ability to do otherwise in some metaphysically robust sense, and if the world is not like that, then everyday judgments of blame will be in error. Call this an *essential error*.

Standard philosophical accounts of blame have tended to be sensitive to the possibility of both token and essential errors. However, there is a comparatively neglected third class of error we might think of as systemic or *collective*. We, as a community, might blame everyone too much or too little because, collectively, our blame is, from the standpoint of morality, miscalibrated. Alternatively, our blame might be mistargeted or wrongly selective, focusing on or exempting people of particular social identity groups in ways that diverge from what is justified or normatively permissible. Collective errors are realized in individual token errors. What distinguishes them from typical token errors, though, is their reflecting some wider collective social presumption, disposition, or pattern. Because norms of blame are rarely a matter of merely individual commitment, the risk of collective error is an endemic feature of individual and collective blame, and the practices that depend on it.[1]

[1] This is not to take a stand on the question of whether collective phenomena entirely reduce to individual psychologies or not. Even if we conclude that all collective phenomena must be realized in or operate through individual psychologies, some things have a recognizably

This is an essay focused on both individual and collective errors under oppression, especially in connection with retributive attitudes and retributive punishment.[2] The animating thought is that understanding a particular functional feature of blame—a broadly normative functional feature—helps us understand some important ways blame and retributive punishment can go wrong at both the token and collective level. Understanding this normative function tells us something both about the constraints on individual moral blaming and a distinctive class of collective (especially, systemic) risks for practices that depend on moral blameworthiness.

To make this case, I propose a theory of blame—the Mediation Theory—that holds that (1) desert-entailing blame mediates or constrains a range of powerful psychological phenomena; and (2) it does so in light of social and normative interests, including those that underpin punishment practices. Jointly, these features give rise to (3) an important moral hazard for the design of practices of institutionalized retributive punishment—the erasure of the mediation function of blame. This moral hazard is an especially significant one under conditions of oppression. The upshot is this: although an interpersonal blame practice is an important part of our social and moral toolkit, we have reason to think that under ordinary conditions, central statuses in that practice (e.g., blameworthiness) can only support institution-level practices of retributive punishment only if they are very carefully designed to avoid collective errors in novel ways.

The order of presentation is as follows. First, I consider and reject the view that blame is a species of punishment. Second, I reconsider the idea of blame as a kind of reactive attitude, and I provide a novel way of understanding that idea. Then, drawing from work on psychological accounts of the origin and ongoing function of punishment, I propose a new account of blame, the Meditation Theory blame. Finally, I show how the account casts light on the limits and risks of retributive punishment under conditions of oppression.

---

collective aspect to them, one readily identified by pointing to its shared or multi-agent nature.

[2] For a sampling of other recent discussions of what I'm calling collective errors in blame, see Ciurria (2020), Webster (2021), Zheng (2021) and several of the essays in Oshana et al. (2018).

1. Blame is not a species of punishment

In the sense operative here, retributive punishment (as opposed to instrumentalist and other notions of punishment) has three features: it is responsive to perceived wrongdoing, it is concerned with proportionate desert, and it is such that the offender's deprivation of some good (typically, well-being, rights, or privileges) is regarded as of non-instrumental value. Siblings endeavor to punish each other for perceived transgressions and a group of friends might respond to a betrayal of collective trust by one of its members by excluding the betrayer from further social opportunities for a period. However, retributive punishment is perhaps most notable in institutional contexts (clubs, organizations, employers), and, especially, in the hands of the state through the criminal law.

At first pass, many philosophical accounts of blame fit with all these thoughts. For example, on one familiar approach, moral blame involves a characteristic set of interpersonal affective reactions, what PF Strawson called *the reactive attitudes*. These attitudes are responsive to how agents treat one another, and they seem especially sensitive to perceptions of controlled wrongdoing.

There is also an overlapping story to be told about the effects or ongoing function of these practices over time in human social practices. While punishment of various sorts can be found in other species, many researchers have thought retributive punishment plays an especially distinctive role in promoting and preserving cooperation in humans (Seymour et al 2007; Cushman 2015). And, blame, like retributive punishment, seems to be a tool for reinforcing and supporting intragroup cooperation (McGeer 2013; Vargas 2021).

These thoughts can suggest a particular view about blame, at least in the fault-imputing or "accountability" sense (as opposed to the merely defect-imputing or "attributability" sense, or other notions that do not seek confrontation—see Shoemaker 2015). The thought, which I reject, is this: accountability blame is just a way we punish one another. When we blame, or if you like, when we express the attitudes characteristic blame, such as indignation and resentment, we are punishing the offending agent.[3] An

[3] Daniel Robinson (2002) claims that "ordinary opinions settle for the conclusion that praise and blame are just a species of reward and punishment" (28). Although it is unclear that

advantage of moral blame is that it is relatively efficient: it allows us to punish without requiring the formal apparatus of institutions of punishment, as in the criminal law or the proceedings of an employer. However, in cases where the harm of wrongdoing is significant enough or falls within a domain over which some institution claims authority, then institutional sanctions can be enacted in addition to, or perhaps instead of, moral blame. If all of this is right, then we can fit the philosophical account of blame into a broader psychological and social account of punishment by seeing fault entailing blame as a species of the more fundamental retributive attitudes characteristic of our species.

For all its attractions, this view—call it the *Punitive View* of blame—seems to fail when we think about a range of real-world cases. Instances of private blame are perhaps the clearest example. An employee might think that her boss is culpable for the poor morale of the unit but might not ever voice this opinion or otherwise give evidence of it. Yet, it seems plausible to say that she blames the boss for the poor morale. Her blame is entirely private, and so long as it is, it seems difficult to identify anything that would count as punishment.

It might seem more promising to consider *expressions* of blame. Some have thought that an essential aspect of blame is its protestive nature (Hieronymi 2004; Talbert 2012; Smith 2013). Yet, even public expressions of protest need not count as instances of punishment, either. When one curses the unknown person who drank the last coffee from the department coffee machine without brewing a fresh pot, one's moral anger is plausibly blame. It is less plausibly punishment, especially if no one is around to hear the protestations.[4] In sum, relatively familiar notions of what philosophers

---

either held the view, Schlick (1939) and Smart (1961) are sometimes thought to have held the view as a matter of philosophical theory. Michael McKenna (2013, p. 132 n.9) cites Christopher Bennett, Joel Feinberg, and Henry Sidgwick as each holding that blame is a species of punishment.

[4] Another reason protest does not seem promising as a species of punishment is that it makes it difficult to fit that account of protest with social and political protests undertaken for forward-looking considerations. That is, some protests aim at bringing about something new, and there may be no collective view about whether there is some individual or group being at fault for that absence prior to the articulation of the new demand. Thanks to Sebastián Figueroa Rubio for this point.

call accountability blame seem distinct from retributive punishment as it figures in the aforementioned broadly naturalistic psychological accounts that seek to account for the origins of retributive attitudes (Seymour et al 2007; Cushman 2015).

Perhaps the Punitive View of blame can be rescued. Still, these considerations give us some reason to consider whether there are more appealing ways to understand the relationship between blame, retributive impulses, and institutions of punishment might be available to us. That is the project of this paper.

2. The reactive attitudes and blame

We respond to what others do, and how they seem to us, in a wide variety of ways. Moral blame, of the sort traditionally implicated in ascriptions of responsibility and the holding of one another to account, has struck many as intimately connected to—or perhaps even constituted, by—some subset of interpersonal reactive attitudes (Strawson, 1962; Watson, 1987; Wallace, 1994; Wolf, 2011; McKenna, 2012; Vargas, 2013). Whether these attitudes constitute, or only express blame (understood as a further thing) is not a matter we need to decide at this stage, although my initial focus is on the attitudes themselves.[5] These attitudes are typically affect-laden, and in the negative cases (my focus here), they paradigmatically include resentment, indignation, or anger.

There is a tradition of distinguishing the blame-relevant reactive attitudes by their targets, distinguishing self-reactive (first personal), interpersonally reactive (second personal), and attitudes that are vicariously analogous (i.e., third personal, or at any rate, had in response to third party phenomena). In what follows, I will offer a different regimentation of these attitudes i.e., in terms of the moral content or valence implicated in the attitude. To get a sense of their character, it is helpful to start with some concrete examples. (The focus here is on the interpersonal reactive attitudes, but the basic categories persist across self-reactive attitudes and the vicarious

[5] One difficulty here is that the English-language term 'blame' is multiply ambiguous, referring to particular kinds of attitudes (blaming attitudes), the disposition to express those attitudes (a blaming stance), the expressions of those attitudes (blaming), and the roughly practice-like collection of normative statuses, attitudes, stances, and expressions of those statuses and attitudes that might be instantiated in a time or place (a given blame practice) or the nature of that practice qua blame practice (the "blame system").

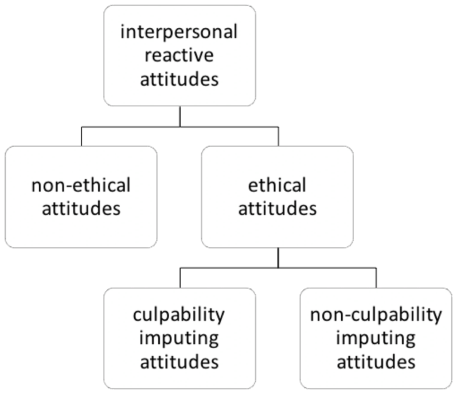
analogs.)

Imagine you and I go to an outdoor festival, and we witness dancers performing in some traditional cultural garb. I tell you that I find the garb and dancing aesthetically unappealing, even repellant, in its own way. You smile at my limited ability to appreciate art across cultures. You note that you think the dancing is beautiful, if not the sort of thing you would take up yourself. After a while, we move on. In a different area of the festival, we come across a group of pre-teens cursing at each other and insulting one another while jostling about who gets to take a turn at a gaming booth. We both shake our head at the bad manners and poor conduct displayed by the children. You note that their obnoxiousness is probably not really their fault, but a product of poor parenting. I reply that I am less confident that this is so, but I concede that their obnoxiousness, self-centeredness, and more global insensitivity to the interests of others may indeed not be their fault. We set out in search of ice cream, finding a vendor on the verge of closing for the afternoon. We each buy a cone and turn to resume our walk. Just then, a thirty-something man wearing a backwards baseball cap comes bursting through the crowd and runs between us to get to the ice cream stand before it closes. In his rush, he knocks both of us to the side, causing you to drop your ice cream and me to spill mine on myself. He looks back briefly to see what has happened, pauses, and then turns back to the person at the register to order himself an ice cream.

In these three encounters, we find distinct kinds of evaluative reactions to the behavior of others. In the case of the traditionally garbed performers, our differing reactions are primarily aesthetic. I find their dancing unappealing, and you find it beautiful. In our observations of the children, we both agree that they are ill-mannered, and perhaps, that they display poor character in their interactions with one another. We disagree about whether that poor character is something for which they are at fault. However, there is a kind of aretaic evaluation about which we agree—the children are, we think, obnoxiously quarrelsome. Finally, in the spilled ice cream case, we unanimously have clear reactions of moral anger. Our anger might take the form of protest, demands for apology or recompense, and so on. However it goes, it is clear that we hold the offender responsible.

The last two cases—the children, and the spilled ice cream—are importantly different from the first. Unlike the aesthetic case, they involve what we might think of as broadly ethical attitudes. However, the attitudes

in the ethical cases seem different from one another as well. They appear to bifurcate into attitudes that are culpability-bearing and those that are not. One way to illustrate the relationship of these classes of reactive attitudes is as follows.



Sometimes the non-ethical interpersonal reactions are less reactions to an action or doing than they are to a being or a seeming. One might involuntarily recoil at another's face, whether by its bare appearance or because of the way it recalls the face of another. Such recoiling need not entail the conviction that there is, indeed, something morally or ethically significant about the person (or action) that is the reaction's proximal trigger. Triggering negatively-valenced interpersonal reactions—including sadness, disgust, resignation, frustration, and disappointment—need not imply an ethical or moral assessment of the person that elicits such reactions.[6]

In contrast, the class of ethical reactive attitudes are distinguished by the perception of moral or ethical significance, whether in terms of wrongful action or in terms of perceived defects in the person. The conduct of the children matters, ethically speaking. While the stakes are relatively low in the case of spilled ice cream, the moral offense is relatively clear. Yet these ethical

[6] To be sure, there may be something morally significant in our having non-ethical interpersonal reactions. For example, if we find that we systematically have apparently non-ethical reactions to people of specific groups, simply in virtue of their membership in that group, we have reason to consider whether we are unduly prejudiced. The point that matters here is that our interpersonal reactions, whether positive or negative, do not always present themselves as moralized reactions.

reactive attitudes need not involve direct interpersonal confrontation. They shape our lives in subtler ways, too. For example, we shake our heads in disappointment at the colleague who, having once again drank too much, indiscriminately talks over everyone else in the conversation. We may privately cringe at the person who, no matter the obviousness of her achievements, seems chronically afflicted with excessive self-doubt.

Some cases with ethical import will be difficult to parse on the matter of culpability, especially where behavior seems to reflect dispositions that seem mostly detached from the effects of deliberative, choice-making agency. Perhaps this is true of both the drunk boor and the chronic self-doubter. Yet many other cases are comparatively clear cut. In the ice cream spilling case, it is prima facie wrongful for someone to shove and thereby injure you. Were someone to do so, this would plausibly trigger your anger, resentment, or indignation. In this mode, these attitudes are culpability-imputing. They depend on the sense that the offender could have better complied with the moral demands, and that his or her failure to do so reflects a defect as a moral agent. Even if there may be forms of moral anger that aren't culpability-imputing, in cases like these, there is relatively clear imputation of culpability.

Yet interpersonal reactions, even to wrongdoing, are not always culpability-imputing. For example, suppose some other case of pushing is an instance where the putatively offending person was saving your life. Alternately, suppose the shove was the accidental byproduct of an unexpected seizure.[7] We might persist in thinking that there are moral stakes here, and that it would have been preferable for you not to have suffered an injury. Yet, anger and resentment would be out of place. Even if we (rightly) felt gratitude towards the person who shoved us, this is not a denial of the shove or the injury, and our reaction to it would not be one of fault-finding.

The foregoing thoughts can suggest either of two different conclusions. First, one might think that it suggests that many of our moralized or ethical reactive attitudes start with the presumption of culpability but then can be defused or mitigated in light of further

[7] My focus here is on cases of non-derivative responsibility. There are cases that are normally understood as instances where agents lack knowledge but where this is their fault (negligence) or where they now lack the relevant control because of some prior decision (e.g., in the case of some forms of intoxication). Such cases are typically accounted for as instances of derivative responsibility, parasitic on some prior instance of responsibility.

information.[8] That seems right, so far as it goes. Second, one might think something stronger, i.e., that all ethical reactive attitudes are at least initially culpability-imputing.

There is reason to doubt this stronger claim, that all ethical attitudes are initially culpability-imputing. Consider the disagreement that figured in the case of the obnoxious children, above. One can think someone is given to cruelty or that they are unkind, while simultaneously thinking that the considered person is not to blame for having those traits. Inasmuch as our reaction is simply to the person having that trait, my reaction is an ethical one, but not a culpability-imputing one. (This is perhaps one way to parse concerns about attributability of the sort that figures in the work of Watson (1996) and Shoemaker (2015).) The important point here is that I might also think that acting on those traits in a given case is a kind of wrong for which the agent is culpable. For example, if it was suitably in the agent's control whether he or she acted on those dispositions, then I might have both the aretaic and the culpability-imputing reactions. One way to understand disagreements over the moral significance of the behavior of older children just is as a disagreement about whether there was suitable control (or what have you) to underpin the specifically culpability-imputing reactive attitudes.

Following Strawson, we might say that the particular subset of culpability-imputing attitudes seem to reflect perceptions of quality of will, or expectations about how our interests matter in the choices of others, and the judgment that the offending agent has not met that standard.[9] For a reactive attitude to be culpability-imputing, it must assume that the agent had the ability to act as morality required (Kelly, 2013, p. 245). This seems to imply a deontic unity to culpability-imputing judgments, turning on a presumption that agents ought to comply with moral considerations in a particular way, and that failures to do so should be met in a condemnatory

[8] I take it that Figueroa Rubio's (forthcoming) account of a defeasible presumption of voluntariness in human behavior has an analogous structure, which suggests a relatively wide-ranging set of presumptions we take as defaults in interpreting human behavior.

[9] Alternative construals of this unifying thread might include the thought that these attitudes signify criticizable attitudes on the part of the agent to whom we are reacting, or that such attitudes are reflections of assessments about interpersonal relationships or their suitability, although the latter may not have the resources to support a system of criminal responsibility.

way (Darwall, 2006).

In contrast, non-culpability imputing reactions contain no such necessary implication. These reactions can be in response to aretaic or characterological considerations, or to some exculpatory transformation of a prior culpability-imputing assessment. We can think, for example, that a person may be disposed to cowardice or neurosis or generosity through no fault of his own. We might also accept that it would be a good thing if people were less cowardly or neurotic. However, in reacting to someone's perceived cowardice (or what have you), we can—but importantly, need not—conclude that the person is culpable for that trait. These interpersonal reactions can still reflect a variety of normative interests, but they stand apart from the distinctive class of culpability-bearing reactions.

The foregoing discussion suggests a view about one thing blame is: it is a reactive attitude of the culpability imputing variety. This is too tidy, though. One way the English language term 'blame' is ambiguous is that it picks out different classes of attitudes, one more cognitive and one more affective (Vargas, 2013; Vargas 2021). The first kind of attitude is the judgment-like attitude mentioned above, what we can think of as blame judgments, or perhaps better, judgments of blameworthiness. Blame judgments, or judgments of blameworthiness, hold that the offender deserves anger, condemnation, ostracism, or so on.[10] The second kind of attitude is the affect-laden reaction, what we might call blaming reactions. Blame reactions include the experience of characteristic affect-laden attitudes such as anger, resentment, and indignation.

Notoriously, blame judgments and blaming reactions do not operate in lockstep. One might judge that someone is not blameworthy, but still feel resentment and indignation towards the considered offender. Although one might be tempted to think that in such situations, the blaming reactions are inapt, one might instead think that the persistence of the attitudes suggests that the blame judgments were wrongly decided. In conflicts between the

[10] Perhaps there are cases where it is permissible to blame someone without thinking that, for example, it is a non-instrumental good that the person be blamed. If so, then either retributive blame is only one kind of blame, or else such cases are not really blame (perhaps they are more pedagogy, or blame-like manipulations), or they are liminal or "twilight" cases of blame, or they are blame cases whose desert elements are "masked" by some countervailing normative pressure that frustrates the ordinary non-instrumental value of blame.

head and the heart, there are at least two ways forward. Still, this suggests a natural solution to the puzzle of dispassionate blame, upon which traditional reactive attitude accounts (i.e., those restricted to affect) have been thought to founder (Sher 2006, p. 88; Smith 2013, p. 32). Dispassionate blame cases are blame judgments without the characteristic affect. That is, if we expand the class of interpersonal reactive attitudes to include dispassionate, mostly cognitive mental attitudes, then dispassionate blame is still a reactive attitude even though it is not an affective one.[11]

So, we get an initial verdict: on the present approach, blame attitudes are a culpability-imputing response to the exercise of agency in others. Accountability blam*ing* involves the having or expression of attitudes that imply deserved censure or condemnation. This is true in both the cognitive and affective versions of blaming attitudes, and it is what makes affective attitudes without the corresponding judgment-like attitude objectionable. The target of blame reactions without blame judgment can rightly protest that one is being regarded as culpable even though the blamer does not actually judge that the blamed is indeed blameworthy.

Before turning to punishment, it is worth acknowledging that 'blame' is ambiguous in still other ways. 'Blame' can also refer to *expression* of the attitude (blaming), and the behaviors characteristic of that expression (e.g., avoidance behavior, retractions of interpersonal warmth, calls for censure, or finely tuned things like elaborate performances of obsequious behavior in response to unreasonable demands). 'Blame' is also used to refer to the stance

[11] One might try to rehabilitate the Punitive View in light of these resources, but the view remains fundamentally unappealing. It is implausible that we always seek the suffering of others when we blame them (in the sense of having blaming reactions directed at them). Expressing indignation at one's child putting another at risk, is not necessarily to seek the suffering of one's child. In expressing indignation that a colleague has once again left the coffee pot empty, we need not seek to make that person suffer. (We might not even know who is at fault.) Instead, our exasperation can serve to release some frustration, even if all we hope for is that the coffee pot be refilled when emptied. Moreover, there are cases where one might express some blaming reaction even while thinking the offender's suffering is precisely what ought to be avoided. Imagine that we believe someone has already suffered enormously for some significant transgression. We might not desire to see the offender suffer further. However, avoiding the offender, withholding interpersonal warmth, and displaying other forms of blaming reactions would not be inapt. The scope of our condemnation and the ambit of our blaming reactions is much broader than punishment. The Punitive View cannot readily accommodate these facts.

or disposition to have or express such attitudes even when the attitudes are not occurrent or active in the blamer (e.g., I might still blame someone for a transgression without having thought of the blamed or the transgression in some time). Finally, philosophical accounts of blame can also seek to say something about the general structure of the collection of normative statuses (e.g., blameworthiness), attitudes, stances and dispositions, and the expressions of these things that jointly constitute a blame practice. In what follows, we will begin to build on a story that starts with the attitude but that comes to say something about the distinctive normative pressures that emerge for practices that involve these attitudes.

3. Retribution and the adaptive function of punishment

We now turn from a philosophical account of the nature of blame (qua attitude) to an account of the psychological basis of punishment. Whether there is a normatively satisfying justification for enacting retributive attitudes and practices is a matter for a normative theory of punishment. The proximal goal here is to see what can be learned about the moral psychology of retribution and blame in their own terms. Doing so requires some attention to different ways in which the term 'retribution' figures in different kinds of explanatory accounts. In the biological and psychological literatures, retribution is typically understood as any backward-looking punitive reaction to a norm violation. In the philosophical and jurisprudential literature, retribution has a more variegated family of characterizations, sometimes focused on a distinctive class of attitude and other times a kind of justification, frequently for punishment. This difference is not unbridgeable; we will start with a version of the former notion and work our way towards a blame-conditioned notion of retribution of the sort that figured at the outset (i.e., desert-entailing punishment where what is deserved has some non-instrumental value).

Consider the following account of the psychology of retributive punishment. In human beings, retributive punishment is a product of retributive attitudes, i.e., attitudes that express a kind of anger directed at other agents, in light of perceived wrongdoing. Practices of retributive punishment, in both interpersonal cases and in state-sponsored institutions, is a product of retributive attitudes. Retributive attitudes have a distinctive psychological profile. They are backward-looking reactions to past transgressions. These attitudes demand the suffering (or hard treatment,

including privation) of the offender, in response to that wrongdoing. The offender's suffering is viewed as apt, deserved, or perhaps, as an intrinsic good.

These attitudes do not operate willy-nilly. As they tend to exist today, at least among those with Western moral sensibilities, features of the agent and the perceived moral quality of the act can affect our disposition to experience the retributive attitudes. For example, a person who, at the time of the offense, did not understand what he was doing might not deserve punishment. Similarly, a person whose actions did not introduce any improper risk or result in any moral harm, will not tend to elicit retributive attitudes. However, agents who knowingly and intentionally undertake violations of moral norms we regard as binding, especially when there is a clear harm, tend to trigger our retributive attitudes.

Importantly, retributive attitudes are characteristically insensitive to assessments of whether that agent will offend in that way again. The retributive emotions can demand that someone be punished for having harmed our loved ones, even when those loved ones are secure from further harm by that person, or indeed, even when the loved ones might not exist anymore. Thus, retributive attitudes, like many emotions, bind agents to courses of action that might otherwise be undermined by downstream considerations. They are commitment devices: retributive attitudes motivate us to punish even when it is costly for us to see it through.

This can seem extraordinarily puzzling. Why did we become creatures that were so easily disposed to punish others (and ourselves), even at great cost to our own happiness and well-being? Perhaps the standard explanation within the biological and psychological sciences is an adaptive one, tied to the social consequences of backward-looking retributive attitudes (Seymour, Singer, & Dolan, 2007; Carlsmith, Darley, & Robinson, 2002; Cushman, 2015). For our purposes, it does not matter if retributive attitudes and norms are psychological basic adaptations, or instead, stable cultural achievements for creatures with (presumably evolved) psychologies like ours.[12] The

[12] One attractive model is given by Nichols, who claims that "our (narrow) retributive norm was not fashioned ex nihilo, spewing forth from rationality. Instead, our retributive norm is a product of cultural pruning. The unfocused retaliatory norms and practices of our ancestors were reshaped and refined, leaving us with the vestige we have today. But anger was likely a sustaining factor throughout this cultural evolution of punishment norms. Had

important point is that among social agents, retributive attitudes provide a deterrent effect to free riders and others who would exploit cooperative schemes. Additionally, they provide a motivation for cognitively sophisticated creatures to learn, internalize, and promulgate cooperative norms.[13] Given our ability to infer punitive intent, to use language to convey explicit norms of conduct, to anticipate how other agents will respond to the effects of retributive punishment and its threat, and the corresponding pressures of reputation management, retributive attitudes have tremendous payoffs for creatures like us.[14]

Crucially, for the deterrent effect of retribution to function, the involved attitudes must be backward-looking and motivationally robust for those with those attitudes. That is, those attitudes, when they are hard, must be so motivating that the affected agent will act on them even when it is significantly costly to do so. Incentives for strategic norm-breaking would increase if retributive attitudes were not quasi-ballistic, backward-looking, and self-binding. For example, absent the motivational force of retributive attitudes, it might seem more attractive to break an agreement in any circumstance where the cost to the victim of pursuing punishment exceeds the loss. Moreover, if punishment were a purely prospective matter, would-be-offenders could learn to strategically signal an inability to learn from punishment, thereby eluding any cost for wrongdoing (Cushman 2015, 123).

---

our ancestors lacked the propensity for anger at wrongdoers, we would likely not have the retributive norm we do today" (2015, p. 130).

[13] The extent to which norms beyond the norm to punish are substantively cooperative might vary across cases. I take it that the thought is not that retributive attitudes guarantee a wide range of substantively cooperative norms, but that having them fosters the acquisition and retention of those norms in comparison to social groups without retributive attitudes. Thanks to Shawn Wang for raising this issue.

[14] We punish agents in ways that suggest highly nuanced conditions for punishing, including sensitivity to the outcome and the degree of perceived causal control of the agent, and whether there is some justification for the act (Cushman, 2015, p. 120). In the psychological literature there are various candidate explanations for these complexities. Among these are accounts that emphasize the instructive dimension to punishment (Funk, Gollwitzer, & McGeer, 2014; Cushman, 2015), the imposition of fitness costs (Rand & Nowak, 2011), partner choice (Hirschleifer & Rasmusen, 1989), and a "cultural pruning" model whereby a basic retributive impulse is conditioned by culturally specific features (Nichols, 2015).

So, for retributive norms and dispositions to succeed, they must be backward-looking in orientation, and relatively insensitive to forward-looking costs.

Although retributive attitudes earn their keep on roughly consequentialist grounds (Cushman, 2015), they do not have that significance for individual agents.[15] From the standpoint of individual agents, there is nothing characteristically consequentialist about the content and conditions for the retributive attitudes. There is nothing pedagogical in the felt function or aim of retributive attitudes and norms. Nevertheless, their individual and cumulative effect is something very much like that, exerting psychological and social pressure to learn and comply with those norms that have currency among one's society.

As an aside, note that if retributive attitudes are acquired in virtue of their affects, this suggests a debunking argument for at least some incompatibilist views about retribution, of the sort that figure in philosophical disputes about free will and moral responsibility. That is, once we have an adaptive account for why we acquired these attitudes in the first place, there seems to be no special reason to suppose that the conceptions of agency that are supposed to establish a basis for deserving retribution (for example, libertarian agency) must actually exist. On this view, our retributive impulse came first because of its adaptive benefits. The subsequent metaphysics of libertarianism was just a speculative overlay that, perhaps, an ad hoc justification for our retributive impulses. If that's right, one might think that the truth or falsity of metaphysics projected on to these attitudes is mostly irrelevant to the question of whether we should enshrine or suppress the retributive attitudes in our practices. Instead, the more relevant question seems to be what these attitudes do for us, and whether the net benefits outweigh the costs. That is, the issue is practical, not metaphysical.

Returning to the main line of the account under consideration, we acquired backward-looking retributive attitudes in part because of what they enabled in creatures like us, i.e., stable multi-agent norms and the attendant disposition for ready norm-acquisition and compliance. The retributive attitudes gave rise to social practices that expressed those attitudes (e.g., formalized revenge-seeking). Given time, contingency, and certain kinds of

[15] In political philosophy and the philosophy of law, this kind of normative structure (i.e., forward-looking justifications for backward-looking reasons or practices) is associated with the work of Rawls and Hart. Recent discussions of its viability in the context of theories of responsibility can be found in Vargas (2013), McGeer (2015), and Vargas (2022).

social organization, these attitudes and practices plausibly contributed to institutionalized organizations for punishment up to the modern nation state's intricate and varied criminal justice systems. Crucially, though, this isn't just a story about the original function or origin of these practices. It is partly a story about why practices of this sort have tended to persist. That is, they continue to enable a complex of values or goods (i.e., multi-agent cooperation and norm-acquisition and compliance) that we have reason to care about in an ongoing way, whatever the basis was of their initial acquisition.

Institutionalizations of retributive attitudes are a codification of those attitudes, but also, an ostensibly "objective" or impersonal mechanisms for enacting retributive practices. These institutions can come to displace a great deal of the face-to-face and small group dynamics that were historically the most visible outcome of those attitudes. Presumably, no small part of the appeal of this phenomenon is that it displaces a good deal of the individual cost of punishing. Other things being equal, if the state identifies and punishes transgressions, there is a better chance that punishment-seekers will not bear the costs of complaint against socially advantaged offenders. Moreover, state identification and organization of retribution can help forestall the risk of cycles of reciprocal violence that can emerge when victims and offenders disagree about whether some instance of retaliation was proper. If the government identifies, prosecutes, and sets the punishment for the transgressions of the Montagues, the Capulets cannot be directly responsible for having voiced the complaint, for pursuing retribution, and for punishing more than the Montagues think is appropriate.

Going forward, I will take this account of the origin of our retributive attitudes as given. At least in broad strokes, this sort of view is well-motivated by the existing scientific literature and is a product of a wide range of psychological, anthropological, biological, neurological, and game theoretic results (Seymour et al., 2007; Cushman, 2015; Nichols, 2015). In saying this, I do not mean to imply that the origins of punishment are now entirely settled. While that foregoing account of punishment has good empirical credentials, it could still be overturned in whole or in part. Nevertheless, the present account provides a useful basis on which to reflect on how a philosophical picture of the moral psychology of blame might be fit into an empirically credible picture of the psychology of punishment.

4. The Mediation Theory of Blame

It is time to begin assembling the pieces. First, I will argue that a system of blame—a collection of practice-like phenomena including blame attitudes, dispositions to form and express them, and norms regulating them—is functionally distinct from the retributive attitudes and practices, and that a blame system achieves distinctive goods. I will then argue that a blame-conditioned system of punishment practices achieves further goods not gotten by a simple system of retribution. This will then set up the argument of the final section, namely, that there are underappreciated risks for institutions of retributive punishment. The emphasis in this section is mostly at the level of kinds of practices and the general features of those practices—this is about system design (whether and how to have particular attitudes, dispositions, and so on in the practice as opposed to not, or as opposed to different ones), rather than what is happening at the level of individual instances of blaming and punishing.

One distinctive feature of a social system of culpability-imputing blame, as opposed to retributive attitudes, is that it has a wider range of mechanisms and attitudes in its tool kit (Mc Geer, 2013, p. 173). Culpability imputing blame has a wider psychological profile than retribution. It can involve a much wider range of attitudes than just moral anger. It can include feelings of disappointment, feelings of being hurt, a disposition to avoid the offending agent, and a withdrawal of interpersonal warmth. Although retributive attitudes plausibly play some role in the origins of that part of blame that is bound up with moral anger, it does not exhaust the range of attitudes that figure in blame and blaming. At best, retributive attitudes propel some of the reactive attitudes that are part of the larger system of judgments, attitudes, and practices that constitute blame in its contemporary form. To be sure, if we focus on attitudes like indignation and resentment, these can suggest a retributive origin in blaming reactions, and an exhortatory, behavior-modifying constitutive aim. But culpability imputing blame is not always like that. Blame can work with softer hands.

One thing that makes a system of blame (of the culpability-imputing sort) distinctive is that it allows moral criticism to play a more elaborate role in our social life than would be afforded by purely retributive attitudes. The varied forms of blame—resentment, condemnation, and disappointment, but also judgments of blameworthiness, standing, and excuse—provide a complex toolkit for moral critique and the signaling of our own commitments

(McGeer, 2013, p. 182; Shoemaker and Vargas, 2021). In contrast to retributive attitudes, when we blame, we need not think that the offender should suffer some deprivation, or that it would be good for the offender to be so deprived. Nor need we think that we are shaping the behavior of someone. To the extent to which retribution can be morally complex—for example, being sensitive to questions of standing to punish—it goes through a notion of blameworthiness.

Culpability imputing disappointment provides one example of how blame can come apart from the retributive attitudes. Consider the thought that it is the nature of parents and children to disappointment one another. Presumably, in expressing such a thought, the operative idea is not merely that, alas, given the vagaries of life there was no other way things could have worked out, for parent or child. Rather, the thought is precisely that even given the vagaries of life, or perhaps, especially because of the particulars, the offending party could have done better. Such judgments are culpability-entailing, but they need not share the retributive tenor that seems more evident in indignation and resentment; the disappointer's deprivation of some good need not be sought or even regarded as a proper non-instrumental good of blame in judgment, reaction, or expression. Fault is being found here, and a culpability-entailing reaction is evidence of it. However, what is sought might be something like an apology, or a recognition of wrongdoing, or a commitment to improve one's efforts, or even merely some fuller account of why things turned out the way they did. Deprivation of some good is not always on blame's menu, and sometimes unwelcome even as a complementary digestif (Duff, 2015; McKenna, 2012).

Moreover, blame can operate even in contexts where there is no hope of it affecting its target. Even when uptake strikes us as unlikely or impossible, we can still blame to protest, i.e., to call attention to some wrongdoing and voice our opposition to it. Protestive blame does not require uptake or transformation in those we blame. Perhaps we are already aware that our targets are insulated from the effects of our condemnation, whether by social privilege, habit, or skepticism about those who make the complaint. Perhaps they are merely dead or absent. Still, we can blame because the norms governing blameworthiness are, like the retributive attitudes, backward looking. Even when blame does not communicate anything to the offender, and even when there is no hope of uptake, or no hope of inflicting suffering,

the status of being blameworthy is a status internal to a kind of practice.[16]

The social role of blame can be brought out by reflecting on what Michael McKenna (2012) has called the "Responsibility Exchange." The Responsibility Exchange is a model of the paradigmatic social dynamics involved in holding one another responsible. It comes in three stages. The first stage is the offender's morally significant act. The second stage is the reaction among the offended—characteristically, a reactive attitude, and frequently some corresponding practice of holding responsible (e.g., condemnation, social distancing). At the third stage, the offender replies. At this stage, the offender acknowledges the offense, either accepting culpability, perhaps asking for forgiveness, or casting the significance of the action in a different light by disavowing or distancing the offender from the act.

These stages do not always progress in a tidy sequence. Some offenders will simply resist providing a plea or acknowledging offense. Those offended can in turn give offense to the initial offender, starting a new responsibility exchange. And those offended might continue to blame even after some account or apology has been issued. What matters for present purposes is that the dynamics of the Responsibility Exchange highlight the social role of blame.

Blame must function as a mediator between the interests of the offended (which may sometimes be retributive) and the social interests of the offender (who may not have sought to give offense, and who may remain a worthwhile and reliable cooperator). Blaming practices frequently provide—and even invite—the offender and others to provide reasons to modify the various attitudes the offended might adopt. For blame to function as it does, it must be sensitive to push back from the offender, or to information that transforms the significance of the offender's act. Pulling the lens back to blame's wider normative and social function—which need not be present in the minds of blamers and blamed—having a blame practice is one way of mediating between competing individual interests, the larger benefits of social cooperation and coordination (including the formation of moral considerations-sensitive agency), and the threat of retribution by offended agents (Vargas 2021). Even when we privately blame, the fact of

[16] For more on blame as protest, see Angela Smith (2013). For more on doubts about the need for communicative uptake, see Vargas (2016). For discounting of protest, see the literature on epistemic injustice see Fricker (2012).

blameworthiness helps us to organize and coordinate our responses to the offender. Even when we do not express our blame—as in cases where a group may collectively withhold expressions of blame, judging that it is unnecessary or in poor taste—the offender's blameworthiness is still a socially recognizable status that matters to us. Keeping track of this is important, in part of because of what it licenses us and others to do.

On the present account—call it the Mediation Theory—an important and distinctive role for a blame system is that it provides a way of negotiating between our individual impulses of anger and complaint and the varied interests of our complicated social world. Blame provides hope of resolving conflicts before the reactive psychological push becomes a reactive shove by creating a set of propriety conditions around which reasoning, dialogue, and criticism can alter conduct and conviction, gradually shaping our practical dispositions.[17] The degree to which particular reactive attitudes, including those grounded in retribution, play prominent roles in our moral lives is an empirical question. But the socially mediating—and socially mediated—role that blame plays is also reflected in some of the corresponding normative content of our ideas about punishment. From the standpoint of normative concerns, it may seem that retributive punishment (whether interpersonal or institutional) is something that is only justifiable, if it is, after the offender has had some fair opportunity to participate in the responsibility exchange.[18] In the context of state institutions, trials might be viewed as a formalized version of this function.

However, interpersonal contexts of punishment are somewhat more chaotic than institutional models suggest. Sometimes, we do think it makes sense to inflict suffering on people without allowing an opportunity to dispute our assessment of culpability. Some egregious offense might license an effort to punish by, for example, inflicting the "silent treatment" on that party. This silence is retributive to the degree to which it is undertaken to

[17] This picture offers one way of unifying important threads of the broadly communicative picture of blame among some theories of responsibility–as in McKenna (2012) and Fricker (2016)—with the broadly instrumentalist approach put forward by McGeer (2013) and Vargas (2013), among others.

[18] For the notion of fair opportunity, see Brink and Nelkin (2013) and Brink (2021); for more on the responsibility exchange, see McKenna (2012).

make the other party suffer.[19] However, if the underlying cause of blameworthiness can be addressed, or social relations can be re-knit (via apologies, transformation in conduct, restorative efforts, and so on on), then the punishment loses its underlying warrant. Unless the offense is part of a background of systematic wrongdoing, or some especially significant act, it can seem unjust for the offended party to refuse to participate in the Responsibility Exchange. Thus, although interpersonal retributive punishments are at some remove from the operations of state institutions, blame continues to play an important mediating role in both cases.[20]

The appeal of seeing blameworthiness as prior to punishment may be an outgrowth of the mediating function of blame. If retributive reactions simply operated independently of any nuanced assessments of culpability, then blame would be of little use in mediating some of the most potent interpersonal reactions. Blaming reactions (and relatedly, judgments of blameworthiness) are better perform at their mediating role if blameworthiness is a precondition for retribution. It is, of course, a further thing to show that a model where punishment is conditional on blameworthiness (call it the "blame-conditioned" picture) is more normatively appealing than unconditioned, simple retributivism about punishment. We

---

[19] If interpersonal punishment requires that the punisher have authority to inflict punishment, I suspect that such authority is readily present in ordinary adult social contexts.

[20] Moral blame has diverse proximal functions. Among its functions are these: to express moral protest; to morally influence behavior; to elicit from the offender an account of their motivations for acting; to align the moral sensibilities of others with our own; and so on (Wang 2021). It is doubtful that every instance of what we ordinarily recognize as blame has all these features, and for any privileged feature of the proximal function of blame, there will likely be cases that are intuitively instances of blame that do not readily fit this model. For example, the classical utilitarian theory of blame as moral influence foundered on the fact that much of blaming is backward-looking, or for the record, as it were. And blame, understood as an activity bent towards the alignment of moral sensibilities, or as a form of moral address, struggles to account for absent or deceased agents, as well as the phenomenon of private blame. The best versions of these accounts have things they can say about such worries. For my part, I'm inclined to think that we can allow that all these things are functions or aims of blamers when they engage in blame. The account I favor about the chief unifying normative element to blame (see Vargas 2013, 2021) understands the variegated forms and proximal functions of blame to be a set of interlocking attitudes and practices that jointly foster and extend our ability to recognize and respond to moral considerations.

will return to this issue below. However, if we assume that blame-conditioned punishment is appealing, its coherence with several familiar thoughts about the criminal law is immediately apparent.

For example, the idea that blameworthiness is prior to deserved punishment explains some of the appeal of one familiar view about (non-strict) criminal liability, according to which it has a basis in moral culpability (Brink 2021). On this view, a necessary condition on just criminal punishment is that the offender is morally responsible for their transgressions. People who offend without being morally responsible for their transgressions cannot deserve retributive punishment. On the present account, this link between criminal liability and moral blameworthiness is not an accidental feature of local moral convictions. Rather, given that the psycho-social point of blame just is to mediate our interests among agents prepared to punish at considerable cost to themselves, blameworthiness must be a condition of deserved punishment. Otherwise, blame would be entirely ineffective at corralling the scope of retributive attitudes, and thus, ill-suited to creating space for more nuanced moral reactions to perceived wrongdoing. Blame-conditioned retributive punishment retains the cooperation-enhancing effects of punishment, while securing the benefits of a system of blame.

A parallel logic is operative in the *mens rea* requirement. The introduction of a formal mental requirement in legally significant blameworthiness—*mens rea* in the criminal law—was a kind of cultural achievement that limited the scope of punishment.[21] In retrospect, the appeal of a mental element in culpability is evident. People are on the hook for less, and they can better anticipate when they could be punished, and in turn, this allows people to better control the shape of their own lives. The cost of building in mental requirements on culpability is, of course, that we are left with the difficult problem of inferring mental states in offenders. The

[21] Chesney (1939) finds the origins of criminal law as a response to blood feuds, and notes that liability tended to be imposed on the offender quite apart from intent. Anglo-Saxon law from the 1100s was sometimes explicit that "one who does wrong unknowingly must suffer for it knowingly" (cited in Raymond (1936, p. 95). Chesney maintains that *mens rea* became a more systematic part of common law via the influence of canon law, but that prior to that "a criminal intent was not always essential for criminality and many evil doers were convicted on proof of causation and without a proof of an evil intent to harm" (630).

benefits presumably outstripped the costs. Moreover, it seems likely that once such a conception of culpability was in play, it reoriented our moral sensibility, transforming our sense of appropriate or fair ways of treating one another.

According to the Mediation Theory, blame and the retributive attitudes are functionally distinct from one another. Retributive attitudes play an important, if comparatively coarse-grained function in enabling cooperation. Blaming practices does something similar, but such practices mediate the force of retributive attitudes (among others) in light of wide-ranging demands on social cooperation and coordination.[22] Both blame and retributive attitudes are given to some degree of local cultural loading about their conditions. In some places, it is fair game to punish the kin of the offender. In some times and places, we may have done without anything like a mens rea requirement for blame and culpability. Yet in its current form, blame has a much more complicated interpersonal profile that the retributive attitudes, and this complexity shapes our moral lives in ways that would have been impossible with retribution alone.

It is likely that the particular shape of blame in the contemporary world is a consequence of overlaying a culpability requirement (i.e., blameworthiness, especially where it is understood to include a mental element) on expressions of retributive attitudes. Some distinctively modern and culturally local outgrowths of this way of corralling retributive attitudes may include the idea that blameworthiness comes in degrees, the persistence of disputes about just what the mental elements come to, the proliferation of practical strategies for dealing with the opacity of mental states, and the evident complex judgments about attempts, negligence, and the phenomenon of moral luck. The complexity of our blaming practices is, in part, a reflection of complex pressures of taking seriously a mental requirement on our moral anger while balancing various social interests,

[22] It is an empirical matter whether, from the standpoint of facilitating social cooperation, a blame-and-retribution system always does better than an exclusively retribution-based system. My suspicion is that even if retribution-only systems are better than alternatives in small bands, retribution's destabilizing effects tend to outweigh its cooperation-enhancing features in more complicated forms of social organization. However, it may be that in almost all forms of human organization, a blame-and-retribution system does better than an exclusively retribution-based system of norms.

including what is signaled by our attitudes, judgments, and practices. In turn, some of the complexity of punishment looks straightforwardly parasitic on these elements.[23]

We will return to the issue of institutional punishments in a moment, but before doing so, it may be fruitful to remark on the normative virtues of this picture of blame.

5. Normative authority

It is one thing to give a broadly naturalistic story about the practice of blame and how it functions. It is another thing to show that it can have normative authority, that there is good reason for us to want, endorse, or accept it as a systematic social practice. The account thus far has been mostly explanatory of our moral psychology and practices. However, there is a normatively attractive picture lurking in the Mediation Theory, so it is worth drawing out its basic shape.

Recall the distinction suggested above, between a picture of retributive punishment conditional on blameworthiness, and an unconditioned, or simple picture of retributive punishment. From a functional standpoint, both are systems of social regulation that enable stable practices of cooperation and coordination. However, the present account suggests important reasons to favor a blame-conditioned picture of retribution.

One advantage of a blame-conditioned system of punishment is that is comparatively more sensitive to a broader range of our interests, and scales back our vulnerability to punishment to a range of phenomena over which we have more control, in some relevant and recognizable sense.[24] A system of

[23] Can blaming practices detach from retributive attitudes? Perhaps. The picture on offer here holds that blame in its most familiar current form is rooted in refinements of retributive attitudes via conditions superadded by culture or normative pressures. Complete detachment from our retributive psychology might be possible even if our retributive underpinnings don't disappear. For a thoughtful discussion of this issue, see Pereboom (2021); for reservations, grounded in the contribution of responsibility practices to social coordination and cooperation, see Vargas (2021).

[24] There are a variety of accounts that offer ways to understand the responsibility-relevant notion of control, including, among others, Fischer and Ravizza (1998); Vargas (2013, p. 209-238), and the fair opportunities accounts in Brink and Nelkin (2013) and Brink (2021).

blame, built around a notion of blameworthiness, readily allows for protestive pushback from blamed parties. It also allows the effects of blame to weigh in the blaming, and the assessments of the community at large. These features enable a wide range of opportunities for repairing interpersonal relations. Blame does this, in part, by restraining the terrible anger of retributive impulses to cases of blameworthiness, and the expectation of the Responsibility Exchange creates a dynamic of bi-directional negotiation of moral repair. So, there are plausibly distinct goods that are not present in a model of retributive punishment unconditioned by blameworthiness.

For those of us who, by cultural acquisition or philosophical endeavor, are committed to a blame-mediated picture of retribution, the idea that criminal punishment should be retributive can seem particularly potent.[25] Absent a concern for desert, wrongdoing, and retribution, the criminal law looks like a shoddy tool for social engineering. If crime is only something to be reduced or incentivized, it is not clear why criminal punishments are a very good way to do those things, as opposed to pursuit of social policies that more directly target the sources of crime. By linking punishment to culpable wrongdoing—and thus, to desert—a retributive system promises a system of criminal justice, as opposed to a system of ineffective social manipulation. So, the desert-based, blame-mediated picture of retributive criminal justice seems to both capture important social goods, and at the same time, explain the animating logic of the criminal law.[26]

---

[25] Beyond its effects, a blame-conditioned model also captures a web of important ideas in our moral practices, and in the criminal law. As David Brink (2012) has noted, "Morality, as well as criminal law doctrine, distinguishes between two ways of avoiding blame—justifying and excusing conduct. Justification denies wrongdoing, and excuse denies responsibility for wrongdoing. Insofar as moral retributivism says that moral blame ought to track desert, where desert is the product of the two independent variables of wrongdoing and responsibility, it fits our moral defenses like a glove" (500). For us, the standard desert basis for moral and legal retributive punishment goes through moral responsibility.

[26] One might worry that when we consider existing institutions of punishment, it is easy to identify punishments that do not comport with a picture according to which retributive punishment presumes blameworthiness. Consider when a judge issues a fine for going 27 miles per hour in a 25 mile an hour zone, even when traffic and pedestrians were absent during the offense. In such cases, it is not obvious that something morally blameworthy is required for punishment. Similar questions arise for strict liability elements in the criminal law, which are insensitive to questions of culpability. In reply, existing criminal law is

When we see blame in the way suggested by the Mediation Theory—as a socially nuanced set of judgments, attitudes, and practices that balances reactive attitudes, individual interests, and the pressures for cooperation and coordination—blame can be seen for the deeply important and normatively attractive practice that it is. To pursue retribution in a way that bypasses blame comes at tremendous cost to clear moral goods. So, there is considerable moral pressure to pursue punishment in a blame-conditioned way, rather than in a way that dispenses with a requirement of blameworthiness. Even if we could somehow abandon a blame-conditioned view of retributive punishment for the unconditioned picture of retributive punishment, it is not clear what that basis would be for such a choice.

6. Institutions of punishment

Above, I noted that institutionalization of punishment was, in many ways, an important innovation in our social toolkit. Institutionalized punishment suppresses cycles of reciprocal violence invited by individual pursuit of punishment. It also reduces the disincentives for the socially disadvantaged to pursue retribution, and just institutions can provide a moral equal footing for addressing grievances. Institutional retributive punishment can be undertaken dispassionately, as an expression of shared values or solidarity with those who have been victimized. So, if retribution is a good, or something we regard as normatively desirable, there are powerful reasons for us to endorse institutions of retributive punishment.

Despite the appeal of institutionalized punishment, there are underappreciated reasons to worry that in a wide range of conditions, institutions of retributive punishment are at special risk of losing some of the goods of both retribution in general, and blame-mediated retribution in particular. Some of the very same features that make institutionalization of retribution appealing tend to cut against the value of blaming.

Recall that the mediating power of blame depends in part on the thought that norm violations can be met in a variety of ways and that grievances can be acknowledged or recast. Importantly, seeing the costs on in-group members of overly punitive blaming can sometimes attenuate our collective enthusiasm for blame, and over time, can shape our individual and

---

plausibly both normatively suboptimal and subject to pressures that are not exclusively retributivist.

collective moral sensibilities. In contrast, institutions of retributive punishment will tend to (1) detach the punishment from the agency of individuals and the community, (2) suppress reflection on the effects of punishment, and (3) displace the social and interpersonal negotiation of moral repair.

First, an effect of institutions of punishment, qua institutions, is that they tend to obscure the effects of blame and punishment as exercises of our own individual and collective punishing agency. When an external agent does the punishing, the anguish and suffering of the punished looks less like consequences of one's own pursuit of retribution. Punishment is no longer owned by a specific punishment-seeking individual, or even by the community most affected by the wrongdoing.

Second, and relatedly, when punishment is in the hands of a third party, one has less reason to reflect on the consequences of one's retributive thirst. Individually and collectively, we have less reason to consider whether the punishment is suitable if we have outsourced those matters to an independent institution. In contrast, in the case of interpersonal blame and interpersonal punishment, it is much harder to avoid questions of suitability and efforts at moral repair. In the interpersonal case, both the target of our blame, as well as any observers, can relatively easily protest perceived injustice with relative ease.

Third, relative to interpersonal blame and punishment, institutions of retributive punishment tend to be relatively simple in their tools, e.g., incarceration, fines, or community service. These tools tend to displace the more elaborate dance of social estrangement and negotiation. In the ordinary case, moral repair often starts with the tentative and partial restoration of social ties. Only gradually is full rapprochement achieved between offended and offender. These features tend to be lost in institutional contexts of punishment. Moreover, invitations to mercy and forgiveness may be harder to elicit in institutional contexts.

To be sure, retributive punishment need not preclude goals of moral repair. Some restorative processes might well be forms of retributive punishment (Allais, 2012), and some forms of institutional retributive punishment might involve procedures that more closely mimic aspects of the Responsibility Exchange. Indeed, various non-retributive considerations (including rehabilitation, crime reduction, and deterrence) might generally have some role to play in the details or degree of punishment, even within a

thoroughly retributivist institution of punishment (Brink, 2012, p. 503). Nevertheless, in practice the nature of institutions, and the inherent pressure to standard, formal (i.e., impersonal), and efficient processes will tend to usurp the organic processes of moral repair that normally operate in systems of interpersonal blame.

In stressing the fact that institutional punishment undercuts some of the goods ordinarily gotten from interpersonal blame and punishment, my point is not that institutional punishment is necessarily disconnected from the goods of blame and desert. A desert-via-blameworthiness system of retributive punishment just is the institution under consideration. Indeed, for all that has been said, delivering deserved institutionalized punishment may rightly trump the costs to interpersonal forms of blame and punishment. The point here is only that the very nature of institutional retribution, qua institution, will tend to cut against many of the goods afforded by a system of moral blame.[27]

These problems—decreased ownership of punishment, diminished reflection on the costs of punishment, and nuanced moral repair—are exacerbated in a range of too-common social circumstances. Clear in-group/out-group relations, of the sort that tend to emerge under conditions of economic, racial, and other forms of segregation will tend to further suppress some of the social and emotional feedback mechanisms that temper retribution and blame. Where punishment is disproportionately directed at an out-group, especially a low status out0-group, it is less likely that in-group members will be aware of the costs, identify with those that suffer the costs, or otherwise find such costs a cause for concern. Consequently, in any social arrangement where punishment disproportionately falls upon a low-status out-group, we can expect what is plausibly a collective error: retributive attitudes will be comparatively less restrained.

The "ratcheting effect" in criminal punishment displays a similar shortcoming. A ratcheting effect occurs when something is subject to escalation or intensification, but where de-escalation or reduction is unlikely or difficult. One source of institutional ratcheting effects come from the psychology of punishment in one-off cases. In considering a potential

[27] To be sure, if we understand retribution in a different way, as disconnected from moral anger of the sort that figures in the present account of retribution, then these concerns are less urgent.

punishment (e.g., more time, harsher conditions of incarceration, higher fines, and so on), individuals will tend to think about the proposed punishment in light of the horror of the imagined crime. Banning an inmate's access to books, or recreational time, or what have you, will seem an apt response to the offense. So, it will seem puzzling that thieves, rapists, and murders should enjoy such things.

From the standpoint of justice, however, the retributive impulses elicited by such questions tend to distort our appreciation of these issues in institutional contexts. Partly, this is a matter of the relative invisibility of the effects of our punitiveness. The distorting effects of such questions is also a byproduct of relatively pedestrian features of what we attend to when we think about appropriate punishment. From the standpoint of a normatively ideal retributivism, the more miserable prison is made to be, one might think, the shorter or more quickly the amount of deserved suffering is achieved. Yet, such considerations—never mind the larger set of systemic considerations that might plausibly shape a system of criminal punishment, including treating comparable offenses comparably—are less apparent when our attention is directed at something we contemplate some isolated instance of a morally outrageous offense.[28]

The social context of institutions matters a great deal for the risk of ratcheting effects. In countries with low social trust, where sentences can be set by legislation, and where "getting tough on crime" has popular appeal, ratcheting effects are especially likely to emerge (Lappi-Seppälä & Tonry, 2011). In such environments, it is easy for policymakers to signal intolerance for wrongdoing by calling for greater punishment. The consequence is, from the standpoint of justice, an alarming one for an institution of retributive punishment.

In the context of low social trust, the temptation will be to ratchet up the amount of punishment doled out by institutions of retributive punishment. Offenses will seem more outrageous, especially if the offense is by an out-group member, and directed at an in-group member. Again, in socially stratified societies, the costs of escalating punishment will be disproportionately born lower status groups. And again, the individual and

[28] Some have thought that outrage-triggered ratcheting effects are mainly aspirations for increased deterrence. However, empirical work suggests that patterns of intensified punishment is less about deterrence than it is about the expression of moral outrage (Clark et al., 2014).

collective effects of increased punishment will be mostly invisible to those calling for greater punishment. Worse, complaints by members of those doubly disadvantaged groups (i.e., disadvantaged by social group, and disadvantaged by the stigma and costs of criminal punishment) will tend to be subject to discounting precisely because of their subordinate status. Although I will not pursue it here, two other potentially compounding effects are worth noting in passing: (1) the presumption of institutional authority ("well, he wouldn't be on trial or punished by the state so much if he weren't really guilty and deserving") and (2) the possibility of "looping effects" (Hacking, 1995). A looping effect occurs when social roles and their associated scripts or received norms of conduct and disposition help to create the very status under consideration. So, one might worry that in suboptimal social environments, there will be manufactured or prejudicial statuses that tend to create bearers of that status (e.g., "thugs" and "superpredators"), and that the presumption of institutional authority will make it difficult to disrupt the narrative that causes people to see offenders as especially deserving of punishment. In short, conditions of social stratification and oppression will tend to intensify the collective errors afflicting blaming and punishing practices.

If all of this is right, then the implementation of formal institutions of retributive punishment will, in many social contexts, be especially morally fraught. Even if we accept that blame and retributive punishment are normatively appealing in interpersonal contexts, it is far less clear that such benefits travel well to ordinary institutional contexts. In contrast, it is less clear that alternative models of institutional punishment—crime reducing, restorative, rehabilitative, and so on—are as vulnerable to the risks of disproportionate moral anger being directed at stigmatized populations. If our goal is simply to reduce crime, for example, elevated out-group anger might sometimes distort evaluations of what system of penalties is, in fact, effective. However, the risk of distortion is here somewhat more limited by the sense that responding to deserved moral anger is not a central aspiration of punishment.

None of this is to reject the thought that any institution of punishment faces moral risks, simply in virtue of being an institution of punishment. Nor is it to deny that, on balance, the goods afforded by an institution of retributive punishment may exceed the moral hazards I have identified. Indeed, for a range of offenses, it may be that retributive

institutional punishment is the best we can hope for. I take no stand on these issues, and in an already long essay all I can offer is a relatively thin recommendation that we consider whether we might re-introduce some of the elements of individual moral blame that are typically lost in institutional contexts. From the armchair, there is no obviously best way to do this. Efforts in that direction might include reviews of either (or both) individual judgments, or sections of criminal codes, or past sentences in a formalized review procedure. Ideally, this would include and perhaps even emphasize the convictions of the wider communities most affected by them. Alternatively, one might seek to expand efforts to move away from incarceration as punishment, which makes mostly invisible the effects of punishment, as the primary tool of the criminal law. The deprivation of goods entailed by retributive punishments need not be so comprehensive as incarceration tends to be, and it is worth more consideration whether there are ways to achieve the goods of retribution in a more selective way that supports the communicative, dialogic, and restorative possibilities afforded ordinary interpersonal blaming.

Whatever the right positive proposals should be, the argument here has attempted to bring into focus a particular challenge for responsibility and punishment under everyday conditions. Even if one favors retributive punishment in individual and institutional forms, there is a real, seemingly unavoidable hazard here that arises for specifically retributive institutions. Or, to put the concern differently, in contemporary society, the socially nuanced features of blaming practices do not seem to readily scale up in an institutionally satisfying way. Consequently, the goods of blaming—including the suppression of otherwise unhindered (and, sometimes amplified) retributive attitudes—are likely to remain elusive, at least when we pursue retributive punishment in institutional contexts.

Institutionalized retributive punishment is not just an enlargement of the functional features of interpersonal practices of blaming and punishment. Under a range of common social contexts, institutional retributive punishment will tend to operate in a way that is at odds with the some of the morally attractive features that underpin retributive punishment. The most basic conditions of enacting a retributively justified system of punishment requires taking seriously our moral anger while at the same time casting a skeptical eye to its intensity. At the very least, it means that institutional systems of punishment that seek to satisfy retributive ends are

faced with substantial cognitive and moral demands in the balancing of moral hazards and goods. As ever, just punishment remains a difficult thing.[29]

[29] Gregg Caruso, Tom Clark, and Stephen Morris have each pressed me on the relationship of retribution to my views about responsibility. Even though this essay only scratches the surface of the issues they had in mind, it is partly an effort to address their questions. For feedback on this paper, or its ancient ancestors, my thanks to Richard Arneson, Santiago Amaya, Michael Bratman, David Brink, John Doris, Sebastián Figueroa Rubio, Ron Mallon, Per-Erik Milam, Sam Murray, David Shoemaker, Daniel Speak, and Shawn Wang, as well as members of audiences at a meeting of the Moral Psychology Research Group, the University of California San Diego, Dartmouth College, the University of Gothenburg in Sweden, and the Universidad de los Andes.

References

Allais, L. (2012). Restorative Justice, Retributive Justice, and the South African Truth and Reconciliation Commission. *Philosophy and Public Affairs*, 39(4), 331-363.

Brink, D. O. (2012). Retributivism and Legal Moralism. *Ratio Juris*, 25(4), 496-512.

Brink, D. (2021). *Fair Opportunity and Responsibility*. Oxford University Press.

Carlsmith, K. M., Darley, J., & Robinson, P. H. (2002). Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology*, 83(2), 284-299.

Chesney, E. J. (1939). The Concept of Mens Rea in the Criminal Law. *Journal of Criminal Law and Criminology*, 29(5), 627-644.

Ciurria, M. (2020). *An Intersectional Feminist Theory of Moral Responsibility*. Routledge.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. (2014). Free to Punish: A Motivated Account of Free Will Belief. *Journal of Personality and Social Psychology*, 106(4), 501-513.

Cushman, F. (2015). Punishment in Humans: From Intuitions to Institutions. *Philosophy Compass*, 10(2), 117-133.

Darwall, S. L. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, Mass.: Harvard University Press.

Duff, R. A. (2012). What kind of responsibility must criminal law presuppose? In R. Swinburne (Ed.), *Free Will and Modern Science* (pp. 178-199). Oxford, U.K.: Oxford University Press.

Figuero Rubio, S. (forthcoming). Negligence, Agency, and Ascription of Responsibility. *Oxford Studies in Agency and Responsibility*.

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

Fricker, M. (2012). Silence and Institutional Prejudice. In S. Crasnow & A. Superson (Eds.), *Out From the Shadows: Analytical Feminist Contributions to Traditional Philosophy* (pp. 287-306). Oxford, UK: Oxford University Press.

Funk, F., Gollwitzer, M., & McGeer, V. (2014). Get the Message: Punishment is Satisfying if the Transgressor Responds to its Communcative Intent. *Personality and Social Psychology Bulletin*, 40(8), 986-997.

Hacking, I. (1995). The Looping Effects of Human Kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal Cognition: A Multi-Disciplinary Debate* (pp. 351-383).

Hieronymi, P. (2004). The Force and Fairness of Blame. *Philosophical Perspectives*, *18*, 115-148.

Hirschleifer, D., & Rasmusen, E. (1989). Cooperation in a Repeated Prisoners'

Dilemma with Ostracism. *Journal of Economic Behavior and Organization*, 12(1), 87-106.
Kelly, E. I. (2013). What is an excuse? In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 244-262). New York: Oxford University Press.
Lappi-Seppälä, T., & Tonry, M. (2011). Crime, Criminal Justice, and Criminology in the Nordic Countries. *Crime and Justice*, 40(1), 1-32.
McGeer, V. (2013). Civilizing Blame. In J. D. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 162-188). Oxford University Press.
McGeer, V. (2015). Building a Better Theory of Responsibility. *Philosophical Studies*, *172*(10), 2635-2649.
McKenna, M. (2012). *Conversation and Responsibility*. New York: Oxford University Press.
McKenna, M. (2013). Directed Blame and Conversation. In J. D. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 119-140). Oxford: Oxford University Press.
Nichols, S. (2015). *Bound: Essays on Free Will and Responsibility*. New York: Oxford University Press.
Oshana, M., Hutchinson, K., & Mackenzie, C. (Eds.). (2018). *The Social Dimensions of Responsibility*. Oxford University Press.
Pereboom, D. (2021). *Wrongdoing and the Moral Emotions*. Oxford University Press.
Rand, D. G., & Nowak, M. A. (2011). The Evolution of Antisocial Punishment in Optional Public Goods Games. *Nature Communications*, 2, 434.
Raymond, P. E. (1936). The Origin and Rise of Moral Liability in Anglo-Saxon Criminal Law. OREGON LAW REVIEW, 15(2), 93-117.
Robinson, D. N. (2002). *Praise and Blame: Moral Realism and Its Application*. Princeton, N.J.: Princeton University Press.
Scanlon, T. (2008). *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge, MA: Belknap Press of Harvard University Press.
Schlick, M. (1939). When Is A Man Responsible? (D. Rynin, Trans.). In *The Problems of Ethics* (pp. 143-158). New York: Prentice Hall.
Sher, G. (2006). *In Praise of Blame*. Oxford University Press.
Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford University Press.
Shoemaker, D., & Vargas, M. (2021). Moral Torch Fishing: A Signaling Theory of Blame. *Noûs*, *55*(3), 581-602.
Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews Neuroscience*, 8, 300-312.
Smart, J. J. C. (1961). Free Will, Praise, and Blame. *Mind*, 70, 291-306.
Smith, A. (2013). Moral Blame and Moral Protest. In D. J. Coates & N. A. Tognazzini (Eds.), BLAME: ITS NATURE AND NORMS (pp. 27-48).

Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy*, XLVIII, 1-25.
Talbert, M. (2012). Moral Competence, Moral Blame, and Protest. *The Journal of Ethics*, *16*, 89-109.
Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford, U.K.: Oxford University Press.
Vargas, M. (2016). Responsibility and the Limits of Conversation. *Criminal Law and Philosophy*, *10*(2), 221-240.
Vargas, M. (2021). Constitutive Instrumentalism and the Fragility of Responsibility. *The Monist*, *104*(4), 427-442.
Vargas, M. (2022). Instrumentalist Theories of Moral Responsibility. In D. Nelkin & D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (pp. 3-26).
Wallace, R. J. (1994). *Responsibility and the Moral Sentiments.* Cambridge, MA: Harvard University Press.
Wang, S. T. (2021). The Communication Argument and the Pluralist Challenge. *Canadian Journal of Philosophy*, *51*(5), 384-399.
Watson, G. (1987). Responsibility and the Limits of Evil. In F. D. Schoeman (Ed.), *Responsibility, Character, and the Emotions* (pp. 256-286). New York: Cambridge.
Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics*, *24*, 227-248.
Webster, A. K. (2021). Socially Embedded Agency: Lessons from Marginalized Identities. *Oxford Studies in Agency and Responsibility*, *7*, 104-129.
Wolf, S. (2011). Blame, Italian Style. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon* (pp. 332-347).
Zheng, R. (2021). Moral Criticism and Structural Injustice. *Mind*, *130*(518), 503-505.