Review of Derk Pereboom (2021) *Wrongdoing and the Moral Emotions*. Oxford University Press, 2021. 204pp. Forthcoming in *The Philosophical Review.*

If no one is morally responsible, how should we respond to wrongdoing? Over the past 25 years, Derk Pereboom has grappled with this question with tremendous ingenuity, rigor, and generosity to his interlocutors. That responsibility skepticism is no longer regarded as a merely notional possibility, or the province of a handful of historical figures, is attributable to his efforts. In *Wrongdoing and the Moral Emotions*, Pereboom offers a new and wide-ranging account of what remains when we reject the idea that people are at least sometimes —and in at least one important sense—morally responsible for what they do.

The book's animating idea is that our responsibility practices employ a class of unjustifiable attitudes of moral anger (e.g., resentment and indignation) that are retributive, in that they reflect the presumption that their targets *deserve* to suffer or experience pain. In earlier work, Pereboom held that this kind of blame should be replaced with "moral sadness" at unwarranted wrongdoing. Here he offers a less emotionally detached and more interpersonally assertive reform of our practices, calling for a stance of moral protest against unwarranted wrongdoing. This stance permits expressions of "measured aggression" directed at wrongdoers, a proportionate and controlled defensive "fury" (72).

The first chapter summarizes Pereboom's deservedly influential views about free will and moral responsibility, namely, that we lack free will of the sort that is required for our being morally responsible in the "basic desert" sense. He holds that deliberation and choice earn their keep by providing a solution to the pervasive fact of epistemic openness, even if determinism is true (25). The second and third chapters present his recommended revision in our moral practices. Chapter two presents a conception of blame that does not presuppose what he calls *basic desert*, but instead, is grounded in several broadly instrumentalist or forward-looking considerations (52-3). Chapter three provides an account of the moral psychology of this attitude of measured aggression; Pereboom argues it can support defensive harm without

presupposing basic desert. The remaining chapters explore the ramifications of this account for questions of criminal justice (chapter four), the possibility of forgiveness (five), love and relationships (six), and the role of hope (seven). Despite the complexity of the issues and the scope of the involved literatures, the discussion is consistently accessible.

In what follows, I focus on the book's two overarching theses: *anti-retributivism* and what we might call *anti-angerism*. Anti-retributivism holds that "we can do without retribution, whether it be in justifying our responses to wrongdoing, or in the emotions employed in those responses" (1). Emotions are retributive when they have a presupposition of deserved pain or harm (3). Anti-angerism holds that "moral anger, whether or not it presupposes basic desert, has too prominent a place in our practice of holding morally responsible, and too central a role in many normative accounts of that practice" (3).

The case for anti-retributivism depends on identifying what makes an attitude retributive and on showing that retributive attitudes cannot be vindicated, that is, be shown to be apt, justified, or true. The more demanding the notion of retributivism, the harder it will be to vindicate; the less demanding, the easier it will be to show that retributivism can be vindicated. Pereboom's retributivism explicitly invokes the idea of deserved pain or suffering, even though there are prominent but less demanding characterizations of retributivism (e.g., in David Brink's recent work) that only require the deprivation of certain rights or goods that reflect the nature and gravity of culpable wrongdoing. This latter characterization does not obviously require the relatively strong form of agency on the part of the wrongdoer that Pereboom thinks is required for his version of retributivism. Pereboom says relatively little about why a stronger notion is the right one. The issue is perhaps complicated by his view that the content of our retributive commitments may not be readily discernible from ordinary thought (35). If so, it makes it unclear what considerations favor the more demanding notion of retribution. The non-error-theoretic retributivist might object that Pereboom has stacked the deck against less demanding and more plausible forms of retributivism.

Pereboom's repudiation of retributivism entails that no one can "basically deserve" suffering or pain. Basic desert holds that a wrongdoer would deserve blame or punishment "just by virtue of having performed the action with sensitivity to its moral status, and not, for example, by virtue of consequentialist or contractualist considerations" (2021, 12). However, Pereboom acknowledges that the theoretical options are not just retributive desert or forward-looking blame. There can be accounts of moral responsibility that rely on a non-basic conception of desert. Consider the idea, sometimes associated with Rawls, that social practices can adopt a system of penalties—e.g., fouls and fines—because of the penalty's positive effects for the practice. In some of these instances, the conditions of application of the penalty are a matter of what the offender has done. That is, the propriety conditions on the penalty can be entirely backward-looking, even if justification for having penalties is instrumentalist.

A first puzzle about Pereboom's handling of non-basic desert is dialectical: after rejecting basic desert retributivism and acknowledging that there might be non-basic notions of desert, Pereboom says very little about it, instead focusing on the appeal of "measured aggression." Yet, given that a non-basic notion of desert is not excluded as a possibility, and given that it would involve less radical transformation of our practice, this available but unexplored option haunts his central argument. If there is a non-basic retributive account available, as some are inclined to think there is, it seems no small advantage if it does not entail a radical transformation of moral anger's psychology.

A second puzzle concerns the explanatory stakes of the idea of basic desert itself. Pereboom claims that a test of basic desert is the Kantian thought that a criminal deserves to be punished, even after society ceases (31). Yet, non-basic desert accounts can pass that test. Recall that we can distinguish (1) forward-looking systemic considerations in favor of having a practice with a given structure from (2) internal to the practice, purely backward-looking considerations of the propriety (or application, or truth, or aptness) of first-order desert judgments. For a practice where there are penalties that are deserved only by satisfying (by stipulation) entirely backward-looking conditions, then even if a society dies off leaving a lone criminal, it could still

be true that the criminal deserves to be punished even if no one is there to punish and even if punishing will produce none of the systemic effects that justified having that practice in the first place. Statuses can persist even if the justification for having them disappears.

If that's right, then basic desert seems to combine two importantly different ideas at two different registers: the propriety of an individual ascription of deserved blame, *internal* to a practice; and the external, independent, and *systemic* question of the basis of having a desert base like so (e.g., one justified on contractualist, consequentialist, or other grounds). To the extent to which basic desert's appeal is that it captures the Kantian intuition, it isn't obvious that desert needs to be basic. So, there is an apparent explanatory puzzle about basic desert: why should we accept its entanglement of a view about the basis of individual ascriptions with a view about a separate matter, i.e., the basis of systemic justification?

Turning to the second animating thesis of the book, anti-angerism, Pereboom holds that, irrespective of considerations of basic desert, the practice of holding people responsible "malfunctions in general and crucial respects" because of moral anger (3). He rightly notes that when people are angry, they misrepresent features of the situation, they are prone to defiance and seeking humiliation, and this alienates others and predisposes people to confrontation. Instead, he recommends compassion, and as we have seen, measured aggression.

The apologist for anger should be unmoved. The fact that expressions of some attitude can distort a practice is unconvincing evidence for the claim that we are better off without that type of attitude. When people are in love, they misrepresent features of the situation. Loving attitudes also motivate confrontation, alienation, and humiliation. Even if most instances of blame go badly and most relationships end in heartbreak, it doesn't follow that we rightly settle whether to keep responsibility or love by counting the success cases and subtracting the failure cases. The question is a systemic one, whether we are *overall* better off with the attitude.

A compelling answer requires a clear picture of what roles and functions these play in our overall lives. Love, anger, and the like, are oftentimes the price of admission into certain kinds of practices and particular ways of being.

They are constitutive of other things. They enable other goods that we have overwhelming reason to want. Moreover, anger that is comparatively unmeasured, in being indifferent to whether it secures goods comparable to its cost, has a variety of anticipatory functions, including the shaping of our sensitivities. Attunement to the risk of angry blame is uniquely effective at shaping valuable kinds of character or moral sensibilities, and, as some psychologists think, it is required for effective social norms and the possibility of cooperation and coordination in creatures like us. Such facts can be decisive for keeping moral anger even if it often goes badly. Analogously, if love leads us to heartbreaking debacles and dubious country music, but the alternative was never to have had humans at all, perhaps we should choose country music?

Pereboom also considers the persistent worry that we may simply be psychologically unable to cease having retributive moral anger. As evidence that we can abandon retributive moral anger, he points to changes over time in who we have held responsible and what kinds of punishments are permissible. He also appeals to a 1970s study of a group of Inuits who rarely express anger, which he treats as evidence for the possibility that human psychology is malleable enough to relinquish retributive moral anger.

Again, though, it is hard to see how these considerations bear on the issue. That we have altered our dietary practices over time doesn't mean that we can do without eating; that we no longer blame people for being mentally ill and that we no longer express anger in the ways that we did in the past is entirely compatible with moral anger being a fixed feature of certain kinds of relationships, relationships that perhaps we should not and perhaps cannot want to do without. Nor is it germane that we can find a group of people who do not frequently express angry blame. Not expressing angry blame is not the same thing as not blaming. Nor is the fact that one specific community does not express angry blame evidence that non-blaming can scale up and be sustained outside of that very particular historical, economic, and cultural milieu. Pereboom is surely right to invite us to consider whether there are alternatives to our current moral psychologies and what the trade-offs might be. Having more to say on its behalf might bolster the tenability of the proposal.

This brings us to a set of questions about the positive proposal that we instead adopt controlled, defensive anger without retribution. Is this an injunction to do something we already do in a wider range of cases? Or is it a proposal for adopting a novel attitude that is available to us but not regularly deployed? Inasmuch as it is an argument for something we already do, the empirical evidence Pereboom offers for our ability to engage in non-retributive measured aggression is the testimony of one military officer and some putative examples from combat sports. In those cases, though, it isn't clear that what is at stake is *moral* anger (the examples neither require nor are limited by engagement with a wrongful threat). More clear-cut cases of moral anger—e.g., athletes motivated by an opponent's disrespect, which is not uncommon in sports—suggest something retributive.

A more promising case of measured aggression comes in the form of an anecdote of a confrontation between Teddy Roosevelt and some bullies that is settled by the bullies getting whupped and Roosevelt subsequently inviting them to a beer. But here too, that story doesn't obviously point to a case of non-retributive moral anger; it is just as easily read as an example of how retributive moral anger has proper limits, and the wisdom of reestablishing good will in its wake. Even first-person reports wouldn't be decisive—recall that Pereboom claims the suppositions of an emotion are not readily available to the one experiencing the emotion. To the extent to which Roosevelt-style examples are *moral*, they seem readily read as cases of retributive anger. To the extent to which they are *measured* responses, this too seems a hallmark of normatively appealing forms of retributivism.

As always, Pereboom is inventive, nuanced, and scrupulously responsive to critics of his views. No book this radical in its aims can hope to secure widespread agreement, but it will undoubtedly be a landmark for future discussions of culpability, moral psychology, hope, and the philosophy of punishment.

Manuel Vargas, UC San Diego