1 Moral Responsibility

A fundamental and familiar part of our personal relationships and our everyday moral practices is making judgments about whether a person is morally responsible for their behavior, and, when we judge that they are, holding them responsible for their actions and omissions. We typically think, for instance, that non-human animals, very young children, and those suffering from severe developmental disabilities or dementia do not satisfy the conditions for moral responsibility. On the other hand, when a normal adult human being knowingly does wrong, it is natural to think that they are (absent any excusing conditions) morally responsible for what they did and therefore deserving of certain negative attitudes, judgments, and treatment. Similarly, when someone does something morally right or exemplary, and we judge they are competent, uncoerced, and aware of what they are doing, we feel that they are deserving of praise and reward. Philosophers, however, have long debated whether individuals are ever morally responsible in this sense and whether our common practice of holding individuals responsible and legitimate targets of various desert-based attitudes, judgments, and treatment is ever justified.

This Element introduces and examines the concept of moral responsibility as it is used in contemporary philosophical debates, and explores the justifiability of the moral practices associated with it, including moral praise and blame, retributive punishment, and the reactive attitudes of resentment, indignation, and, more broadly, moral anger. It begins by identifying and discussing several different varieties of responsibility, including *causal responsibility*, *take-charge responsibility*, *role responsibility*, *liability responsibility*, and the kinds of responsibility associated with *attributability*, *answerability*, and *accountability*. It then argues that the kind of moral responsibility that is of central philosophical and practical importance in the free will and moral responsibility debates is best understood in terms of *basic desert*, the idea that the harm of blame and punishment and the benefit of praise and reward are deserved and fundamentally so, and that such backward-looking desert is thus a basic element of morality. We (your authors) deny that human beings are ever morally responsible in this basic desert sense, and we accordingly advocate *skepticism* about basic desert. However, we also contend that forward-looking aspects of our practice of holding morally responsible, those that feature aims such as reconciliation in relationships and moral formation of wrongdoers, are justified. Thus, upon reconsidering moral responsibility, we argue that certain backward-looking aspects should be rejected, and certain forward-looking aspects retained.

In this Element we will examine the arguments for basic desert skepticism. Some skeptics deny that we have basic desert moral responsibility because they

believe it can be shown to be incoherent or impossible. Others maintain that although this sort of responsibility is coherent and possible, nevertheless, our best philosophical and scientific theories provide compelling reasons for denying that we have it. Often basic desert skeptics contend that we lack the control in action, that is, the *free will*, basic desert moral responsibility requires. Accordingly, they are also typically skeptics about an important and controversial sort of free will.[1] What all basic desert skeptics agree on, however, is that adequate justification for grounding basic desert moral responsibility and the practices associated with it – basic desert-invoking praise and blame, punishment and reward – has not been produced (Pereboom 2001, 2014a, 2021b; Levy 2011; Waller 2011, 2014; Caruso 2012, 2018b, 2021b).

Critics tend to focus both on the arguments for skepticism about basic desert and on its practical implications. Some reject the claim that basic desert moral responsibility is incompatible with determinism, and are also concerned that accepting the skeptical position would mistakenly result in people not getting what they deserve. Some argue that rejecting basic desert moral responsibility would have damaging consequences for morality, the law, society, personal relationships, and our sense of meaning in life. They contend, for instance, that it would undermine morality, leave us unable to adequately deal with criminal behavior, increase anti-social conduct, and jeopardize our sense of achievement and purpose. *Optimistic skeptics* respond by arguing that rejecting basic desert moral responsibility would not be destructive in these ways. They argue that threats to morality can be averted, and that prospects for sustaining good personal relationships and our sense of meaning in life can be vindicated. Although retributivism as a justification for criminal punishment would be ruled out, adequate resources for dealing with criminal behavior remain in place.

1.1 Varieties of Responsibility

When philosophers discuss moral responsibility, what they generally have in mind is the kind of responsibility that makes agents justified targets of certain approving or disapproving attitudes, judgment, and treatment. As we've noted,

[1] Historical advocates of skepticism about this sort of free will include Śāntideva (700/1995), Spinoza (1677/1985), Paul d'Holbach (1770), Joseph Priestley (1778/1965), Arthur Schopenhauer (1818/1961), Friedrich Nietzsche (1888/1954). More recent proponents include Galen Strawson (1986, 1994), Ted Honderich (1988), Bruce Waller (1990, 2011, 2015), Michael Slote (1990), Derk Pereboom (1995, 2001, 2014a, 2021b), Saul Smilansky (2000), Daniel Wegner (2002), Gideon Rosen (2004), Joshua Greene and Jonathan Cohen (2004), Benjamin Vilhauer (2004, 2012), Shaun Nichols (2007, 2015), Tamler Sommers (2007, 2012), Brian Leiter (2007), Thomas Nadelhoffer (2011), Neil Levy (2011), Sam Harris (2012), Gregg Caruso (2012, 2021b), 'Trick Slattery (2014), Per-Erik Milam (2016), Robert Sapolsky (2017), Stephen Morris (2018), Elizabeth Shaw (2019), and Farah Focquaert (2019b); for an overview, see Caruso (2018b).

our practice of holding agents morally responsible has backward- and forward-looking aspects. The core backward-looking aspect is desert – in particular, the deserved harm of blame and punishment for wrongdoing, and the deserved benefit of praise and reward for morally exemplary action. A basic conception of desert is backward-looking to its core. That is, basic desert claims are fundamental, and thus not grounded on forward-looking considerations such as reconciliation in relationships and moral formation of wrongdoers.

One concern raised for skepticism about basic desert is that it is too revisionary of our moral responsibility practices. A more conciliatory and conservative position, defended by Daniel Dennett and Manuel Vargas, aims to ground our moral responsibility practice, inclusive of its desert-based justifications, in forward-looking considerations at a higher level. On such accounts, practice-level justifications for blame and punishment invoke considerations of desert, while that desert is not basic because at a higher level the practice is justified by good anticipated consequences, such as deterrence of wrongdoing and moral formation of wrongdoers. Defenders of this approach maintain that our practice of holding agents morally responsible in this non-basic desert sense should be retained for the reason that doing so would have the best overall consequences relative to alternative practices. While such a view avoids certain objections to basic desert, it faces a number of other concerns (see Section 1.3). As an alternative, we will set out in Section 3 our more resolutely forward-looking conception of moral responsibility, grounded in objectives such as moral formation of wrongdoers and reconciliation in relationships impaired by wrongdoing (Pereboom 1995, 2001, 2014a, 2021b; Caruso 2021b; Caruso in Dennett and Caruso 2021).

To begin, it is important that we first differentiate *moral* responsibility, whether backward- or forward-looking, from several other notions of responsibility. In fact, the term "responsibility" is perhaps surprisingly ambiguous and used in a number of different senses. Sometimes, for instance, we use it to simply indicate the cause of something – as when we say that "Hurricane Katrina was responsible for the destruction of New Orleans" or "The fallen tree branch was responsible for the damage to the roof." This sense, known as *causal responsibility*, is assigned when we say that someone or something is responsible for an event or outcome because he, she, or it caused it to occur (Talbert 2016: 8). Inanimate objects (such as the tree branch) and events (like Hurricane Katrina) can be causally responsible in this sense. So too can agents when they play a direct or indirect causal role in bringing about a particular action or outcome. Note, though, that causal responsibility is far from sufficient for moral responsibility – someone or something can cause an event or outcome without deserving moral praise or blame, punishment or reward, for having

done so. This is easiest to see in the case of inanimate objects and events, but it's also true for agents. An infant, for instance, may be causally responsible for ruining your favorite shirt by getting sick on it, and a Parkinson's patient for knocking your cell phone to the ground because of their tremors, even though neither is morally responsible for what they did.

Moral responsibility is also distinct from *take-charge responsibility* (Waller 1990, 2004, 2011, 2014). Skeptic Bruce Waller argues that:

> Just deserts and moral responsibility require a godlike power – the existential power of choosing ourselves, the godlike power of making ourselves from scratch, the divine capacity to be an uncaused cause – that we do not have. (2011: 40)

Yet, Waller maintains,

> you [nevertheless] have take-charge responsibility for your own life, which is a responsibility you deeply value and enjoy exercising. (2011: 108)

Taking responsibility is distinguished from *being morally responsible* in that, if one takes responsibility for a particular outcome it does not follow that one is morally responsible for that outcome. One can take responsibility for many things, from the mundane to the vitally important. For example, one can take responsibility for teaching a course, organizing a conference, or throwing a birthday party. The responsibility taken, however, is very different from the moral responsibility that would justify blame and punishment, praise and reward (Pereboom 2001: xxi; Waller 2011: 105; for an objection, see Smilansky 2012; for a reply, see Caruso 2018b: sect. 1).

A closely related notion is what H.L.A. Hart designates as *role responsibility*. In introducing the idea, Hart points out that:

> Whenever a person occupies a distinctive place or office in a social organization, to which specific duties are attached to provide for the welfare of others or to advance in some specific way the aims or purposes of the organization, he is properly said to be responsible for the performance of these duties, or for doing what is necessary to fulfill them. Such duties are a person's [role] responsibility. (1968: 212)

We can say, then, that role responsibility refers to the various tasks and duties associated with a particular role or job description, such as being a parent, teacher, or firefighter. Such responsibility, however, does not entail moral responsibility. Autonomous machines and AI, for instance, can be responsible in the role sense since they are typically responsible for carrying out various tasks and duties in virtue of the roles for which they are designed. Just as we say that the surgeon, firefighter, or bridge inspector has certain "responsibilities" in

virtue of their role or job description, we can say that an autonomous lunar rover or an AI used to locate and detonate landmines is "responsible" for performing specific tasks and achieving specific aims and purposes. Role responsibility, then, unlike moral responsibility, is to be understood in terms of the various tasks and actions necessary to fulfill a particular role or job description.

Civil liability is also distinct from moral responsibility. Civil liability is a legal obligation that requires a party to pay for damages or to follow other court-enforcements in a lawsuit. In a car crash case, for instance, the injured party can sue the driver and ask for monetary damages. A civil liability is usually a contractual liability or a tort liability. A tort, for instance, is something that occurs when one person's negligence directly causes property or personal damage to another – and it need not be intentional. An example of an unintentional tort would be someone being injured by a faulty product or someone's pet. As general rule, in tort law the financial harm suffered by the victim as a result of a tort is the only issue. Tort law attempts to adjust for harms done by awarding damages to a successful plaintiff who demonstrates that the defendant was the cause of the plaintiff's losses. Criminal law, on the other hand, is concerned with more than such restitution and compensation. It is also concerned with punishing wrongdoers for their criminal acts, not just as an act of restitution but as an expression of the state's disapproval of both the offense and the offender. Skeptics who reject basic desert moral responsibility can retain civil liability because the restitution and compensation of victims can be justified by appealing to the rights of those harmed together with weaker notions of responsibility, including causal responsibility. That is, civil liability need not assume agents are blameworthy and morally responsible in the basic desert sense, only that they are in a weaker sense responsible for some negligent act that caused harm and are therefore responsible, in the civil liability sense, for compensating victims.

In recent decades, philosophers have also drawn distinctions within the concept of moral responsibility. Prominently, some distinguish *attributability*, *answerability*, and *accountability* senses of moral responsibility. The first of these, attributability-responsibility, concerns actions or attitudes being properly attributable to, or reflective of, an agent's self. That is, we are responsible for our actions in the attributability sense when those actions reflect our nature or identity as moral agents, i.e., when they in this sense are attributable to us (Watson 1996; Eshleman 2014). Attributability-responsibility, however, makes no appeal to desert or backward-looking praise and blame, and hence it is distinct from any desert-invoking sense of moral responsibility. A particular action or attitude may be attributable to me in that it reflects on me, on my deep self, and in particular on who I am as an agent in the world, even if I am not deserving of praise and blame, punishment and reward for it. David Shoemaker

(2011, 2015) discusses numerous cases in which agents fail to be morally responsible in the desert-based sense – due to, for example, mental illness, clinical depression, or psychopathy – even though it remains appropriate to attribute various actions, attitudes, and characteristics to them.

Since attributability-responsibility does not depend on desert, it is also consistent with skepticism about basic desert moral responsibility. Consider the views on these issues of the great physicist Albert Einstein. In a 1929 interview in *The Saturday Evening Pos*t, he said: "I do not believe in free will . . . I believe with Schopenhauer: we can do what we wish, but we can only wish what we must." He then went on to add: "My own career was undoubtedly determined, not by my own will but by various factors over which I have no control." He concludes the interview by rejecting the idea that he deserved praise or credit for his scientific achievements: "I claim credit for nothing. Everything is determined, the beginning as well as the end, by forces over which we have no control." While free will and basic desert skeptics may agree with Einstein that he does not deserve praise for his various attributes and accomplishments, they can nevertheless attribute various attributes and accomplishments to him without inconsistency. We can say, for instance, that Einstein was an *exceptionally original thinker* and *extraordinarily intellectually creative* without presupposing that he was basically deserving of praise for any of those attributes (Caruso 2017).

The *answerability* sense of responsibility, defended by T.M. Scanlon (1998), Hilary Bok (1998), and Angela Smith (2012), is also claimed by some skeptics to be consistent with the rejection of basic desert moral responsibility (e.g., Pereboom 2014a, 2021b; Pereboom and Caruso 2018; Caruso 2021b; cf. Jeppsson 2016). In this sense of responsibility, someone is responsible for an action or attitude just in case it is connected to their capacity for evaluative judgment in a way that opens them up, in principle, to demands for justification from others. If an agent is answerability-responsible for a wrongdoing, it means we can legitimately ask him, "Why did you decide to do that?" or "Do you think it was the right thing to do?" Such questions target the deliberative, reasons-tracking aspects of agents. That is, "To be answerable . . . is to be susceptible for assessment of, and response to, the reasons one takes to justify one's actions" (Shoemaker 2011: 623). The sorts of answers the agent gives stand to reveal their reasons for action, what they take to be important, and salient aspects of their moral character, and subsequent responses may take the form of a demand for apology or a request for reform with the aim of modifying the wrongdoer's moral attitudes and dispositions (Shoemaker 2011: 623; Pereboom 2014a, 2021b).

We will have much more to say about this notion of responsibility (see Section 3), but for now we'll add two points about answerability-responsibility. First, it is distinct from, and not coextensive with, attributability-responsibility, since there are

cases where an attitude or action is attributable to an agent without that agent being answerable for it (Shoemaker 2011, 2015). This can occur in cases when emotional commitments operate independently of evaluative reasons. In such circumstances, agents' attitudes or actions reflect on them, on their deep selves, and on who they are as agents in the world, but they would not be answerable for them. Second, an agent's answerability for an action does not entail moral responsibility for it in the basic desert sense. As we'll see in Section 3, there are ways to ground answerability and its demands for justification in forward-looking considerations independent of desert of any type. Answerability-responsibility should thus be understood to be distinct from basic desert moral responsibility.

The final sense of moral responsibility in this tripartite distinction is *accountability*. Accountability-responsibility is the responsibility that philosophers typically have in mind when they debate desert-based moral responsibility, the kind that makes an agent an appropriate target of various desert-based sanctions and rewards, including the reactive attitudes of resentment and indignation, and retributive punishment.

Shoemaker specifies that to hold someone to account for wrongdoing is "precisely to sanction that person, whether it be via the expression of a reactive attitude, public shaming, or something more psychologically or physically damaging" (Shoemaker 2011: 623). Whereas the answerability sense of moral responsibility is open to free will and basic desert skeptics to endorse, the accountability sense, at least as standardly characterized, is not, since it is typically set out as licensing responses such as resentment, indignation, and retribution (Watson 1996; Shoemaker 2011, 2015). We maintain that it is the accountability sense of moral responsibility, specifically when it is conceived as involving such responses as basically deserved, that separates the parties in the debate.

1.2 Desert-Based Moral Responsibility

Moral responsibility applies primarily to actions and omissions. We understand "action" to denote not only intentional bodily movements, but also to purely mental items such as decisions. To say that one is morally responsible for a good or bad action or omission, in the desert-based sense, is to say that one *deserves* to be praised or blamed, rewarded or punished, for that action. This may include the expression of certain attitudes and judgments, like resentment and indignation, or extend upward to more severe forms of retributive punishment. As Randolph Clarke explains:

> If any agent is truly [morally] responsible ... that fact provides us with a specific type of justification for responding in various ways to that agent,

> with reactive attitudes of certain sorts, with praise or blame, with finite rewards or punishments. To be a morally responsible human agent is to be truly deserving of these sorts of responses, and deserving in a way that no agent is who is not responsible. This type of desert has a specific scope and force – one that distinguishes the justification for holding someone responsible from, say, the fairness of a grade given for a performance or any justification provided by consequences. (2005: 21)

This way of conceiving of our practice of holding people morally responsible goes back to P. F. Strawson's (1962) famous account of the *reactive attitudes*. According to Strawson, in the face of interpersonal wrongdoing it is both natural and justified to express certain types of anger; specially, *resentment*, directed toward someone due to a wrong done to oneself, and *indignation*, the vicarious analogue of resentment, directed toward someone because of a wrong done to a third party (Watson 1987; Wallace 1994; McKenna 2012; Shabo 2012; Brink 2021). Resentment and indignation, in Strawson's terminology, qualify as reactive attitudes.

These reactive attitudes are often accompanied by the supposition that its target *deserves* to be the recipient of the expression of such emotions. For instance, in the face of wrongdoing, one might express a kind of anger or blame, one that intentionally causes pain or harm, because it is believed that the wrongdoer deserves such pain or harm. A paradigm example would be when we confront a wrongdoer with expressions of angry blame (Wolf 2011; Fricker 2016; Bagley 2017; Shoemaker 2018; McKenna 2019). Such blame is associated with a certain negative emotional attitude toward the wrongdoer – such as resentment or indignation; and more broadly, moral anger – that goes beyond the mere absence or withdrawal of good will (Wolf 2011: 335). This kind of moral blame is non-trivially painful or harmful, and, when accompanied with a supposition of desert, qualifies as retributive. The positive counterpart of such blame is deserved praise and reward for good behavior.

When philosophers debate moral responsibility, it is typically this kind of desert-based moral responsibility they have in mind. Note, though, that the kind of "desert" operative here could be understood in either a *basic* or *non-basic* sense. In the basic form of desert, an agent deserves the harm or pain of blame or punishment just because he acted wrongly, given that he was aware or should have been aware that the action was wrong. An agent deserves the benefit or pleasure of praise or reward just because she acted in a morally exemplary way, given awareness of the moral status of the act (the account can be extended to omissions). The desert invoked here is basic because these claims about what agents deserve are fundamental in the sense that they are not justified by further considerations such as the good consequences of implementing them, or the