




ARTICLES

# Compatibilism and Truly Minimal Morality

Travis Quigley 

Department of Philosophy, University of Arizona, Tucson, AZ, USA  
Email: [travisquigley@arizona.edu](mailto:travisquigley@arizona.edu)

## Abstract

I formulate a compatibilism that is distinctively responsive to skeptical worries about the justification of punishment and other moral responsibility practices. I begin with an evolutionary story explaining why backward-looking reactive attitudes are “given” in human society. Cooperative society plausibly could not be sustained without such practices. The necessary accountability practices have complex internal standards. These internal standards may fully ground the appropriateness of reactive attitudes. Following a recent analogy, we can similarly hold that there are no external standards for what is funny; the norms of comedy are complex, but funny is funny. However, this is compatible with moral reasons to change the practices themselves, and therefore change what is fitting within them: in the first instance, a moralistic “that’s not funny” is ill-fitting, but “that *shouldn’t* be funny” can be apt. The analogous reformist position prescribes practices constituting the *minimal* responsibility norms necessary for cooperative society.

**Keywords:** compatibilism; determinism; fittingness; Strawson; punishment

## 1. Introduction

Here is a compelling critique of certain moral responsibility practices, especially (but not only) harsh retributive punishment: due to the fact of a deterministic universe, no one is truly in control of what they do; therefore no one can deserve any particular treatment due to what they have done. This argument has long-standing philosophical and popular currency.<sup>1</sup> It is met with arguments for the compatibility of determinism and free will (of the sort relevant for moral responsibility), often arguing that determinism poses no critical threat to moral responsibility practices because the ordinary distinction between controlled and uncontrolled actions is not affected by determinism.<sup>2</sup> This positions incompatibilists as critics of existing moral responsibility practices and compatibilists as their defenders (usually with some qualifications). By contrast, I want to defend a compatibilism that captures the plausible revisionary force of incompatibilist skeptics. Call this *minimal compatibilism*.

<sup>1</sup>For instance, Pereboom (2001, 2021b), Caruso (2021), Waller (2011), Sapolsky (2023), Harris (2012), and Wegner (2002). The argument also goes through if the universe is indeterministic, but in a way that could not ground moral responsibility.

<sup>2</sup>This basic approach is due to Strawson’s “Freedom and Resentment” (1974). I discuss some variations in this large literature below.

This project should be of significant interest to compatibilists, who generally do not mean to affiliate themselves in practice with harsh retributive punishment.<sup>3</sup> But, although there are many different reasons to disavow harsh punishment, the lurking response to any case for humane reform is that bad conditions for prisoners are actually appropriate, even if they cause bad effects such as increased recidivism, because such treatment is what wrongdoers deserve. The most powerful response to *that* is to diminish the force or import of basically deserved harm on the basis of wrongdoing. It would be powerful to combine the compatibilist appeal to the ordinary distinctions of control with the (typically) incompatibilist demand to reform the very basis of our moral responsibility practices.

Here is the structure of minimal compatibilism. Cooperative society requires moral responsibility practices. These practices must warrant holding people accountable for what they have done, and they must be intrinsically motivational for a critical mass of participants. This requires a shared interpersonal network of reactive attitudes, which are emotional tendencies to praise, blame, etc. Such a practice must satisfy the minimal requirements of stable society. Any particular moral responsibility practice will involve complex norms that permit debate and deliberation, but a practice does not question its basic justification. To do so would risk undermining its intrinsic motivational power.

So far, so compatibilist. But these points about moral responsibility practices do not place the practices themselves outside the scope of moral justification: the system does not ask (or make room for) skeptical challenges, but individuals can and do. If the basic point of moral responsibility practices is to ensure cooperative society – I discuss the Strawsonian *bona fides* of this framing below – then the ultimate normative force must lie in the overwhelming value of cooperative society itself. But cooperative society can be achieved in many different ways; there is wide variation across history, the world today, and (surely much moreso) in the accessible possibility space. Perhaps cooperative society requires practices of punishment; but this will not license any punishment beyond what is truly *required*. And surely no actual society operates at this maximally humane limit.<sup>4</sup>

But how can moral responsibility practices be sensitive to criticism, especially radical criticism, if they operate in terms of an internal logic that is not susceptible to skeptical demands for justification? To make the challenge especially sharp, say that moral responsibility is, as David Shoemaker (2017) argues, *response-dependent*. To be response-dependent means that no complete analysis or grounding of a practice is available beyond, roughly, a listing out of what actually happens within the practice. Shoemaker's analogy is humor. It is impossible to account for what is funny without eventually making recourse to what (some) people find to be funny. Shoemaker believes that the same is true for moral responsibility. Clearly determinism is not relevant to a response-dependent account of moral responsibility practices; hardly anyone in fact modulates their reactive attitudes on that basis. But the analogy also illustrates (going now beyond Shoemaker) the possibility of moral reform. The admonishment "that's not funny" does not quite make sense for a joke that is found amusing in the relevant group, but "that shouldn't be funny" does make sense. And many things that used to be

<sup>3</sup>See, for instance, McKenna (2021), who has significant reservations about punishment practices even while endorsing deserved harm for wrongdoers (McKenna 2020).

<sup>4</sup>Although some may be much closer than others. It is customary at this point to mention Sweden.

funny no longer are, at least in part due to purposeful moral reform. (The era of calling people and things “gay” for a laugh now seems very weird.)

I propose a parallel structure in the case of moral responsibility. The fact that someone simply *is* responsible by the lights of a responsibility practice renders certain responses fitting. This is a reason for making those responses, just as a remark’s funniness is a reason to laugh. To say “they aren’t really responsible” because of determinism is a conceptual mistake. But to say “they shouldn’t be responsible” is apt. If there is no compelling rationale for the particular aspect of the moral responsibility practice in question, it is also plausibly correct if the basic desert skeptical argument has any force at all. In that case the practice ought to be reformed. If the practice is changed, then what is fitting within the practice will necessarily change as well. In this way the compatibilist may support all justified moral reforms.<sup>5</sup>

If some measure of retributive punishment remains after all reforms are made, as seems likely for reasons discussed below, then some individuals will still be punished for things they cannot ultimately control. Is this not still an injustice in the eyes of the incompatibilist? On the view here, there is nowhere for this injustice to reside. By the standards of the practice – the ordinary standards of moral discourse – individuals by and large get what they deserve. (Obviously practices malfunction often enough, but not due to determinism.) By the standards of the reformer, an injustice could only be done if some justified reform is left undone. If all justified reforms are adopted (or at least subjects of agreement with the compatibilist), there is no injustice left. One might worry that social stability could instrumentally require punishing the innocent.<sup>6</sup> But this would misunderstand the structure of the view: punishing the innocent could not be fitting in a moral responsibility practice, and reforming the practice to permit punishing the innocent would not be justified.

This kind of “two-level” account is not entirely new in the recent literature. A structurally similar account is advanced by Vargas (2013, 2015).<sup>7</sup> The approach here is distinctive in two ways: first, unlike previous two-level accounts, I do not claim that the second-level justification wards off all claims of the moral reformer. Instead, it only justifies the minimally retributive moral responsibility practice.<sup>8</sup> Second, the first level is not *grounded* in the second level. Our actual moral responsibility practices are, as Strawson would have it, “given.” The truth about moral responsibility is fixed by responsibility practices themselves. This does not block justified reforms to the practices and any resulting changes in what is fitting within them. It could in principle be that we should altogether abandon moral responsibility practices based on reactive attitudes on this basis, but in fact this can and should only be done to a limited extent. And unless and until revisions actually happen, the appropriateness norms of moral responsibility are left entirely untouched.

For many, this approach will still unhappily evoke Strawson’s “optimist” compatibilist, who appeals to the good consequences of holding people responsible for their actions. A very small minority of contemporary philosophers have defended version

<sup>5</sup>The structural point is not committed to a response-dependent account of moral responsibility. The discussion could be adapted to (e.g.) reason-responsiveness or mesh theories. See note 23.

<sup>6</sup>As in Williams’s famous critique of utilitarianism (Smart and Williams 1973).

<sup>7</sup>See the helpful discussion in Pereboom (2021a) – notably, Vargas (2013) does not present his view as explicitly two level, but accepts the characterization in his (2015) discussion.

<sup>8</sup>So the view is not conservative in the sense Pereboom (2021a) rightly criticizes previous two-level views for.

of optimistic or consequentialist readings of Strawson.<sup>9</sup> Section 2 develops a broadly evolutionary account of the reactive attitudes, and then distinguishes that account from other optimist or instrumentalist approaches. Importantly, the evolutionary argument supports the idea that backward-looking reactive attitudes are essential for functional responsibility practices. Section 3 develops this framework in connection with contemporary compatibilist theories. I argue that philosophical compatibilism regarding what moral responsibility practices *are* would only benefit from adopting the broadly instrumentalist argument for what moral responsibility practices *should be*. Section 4 discusses a challenge to the two-level system's stability, and concludes.

## 2. Evolved compatibilism

This section has two goals. First, I set out the basic case that human cooperative society requires responsibility practices involving backward-looking ideas of moral desert. That is, a critical mass of individuals must see an intrinsic reason to hold others accountable for what they have done.<sup>10</sup> This provides the basic normative force for minimal compatibilism. Second, I distinguish this account from the small handful of contemporary instrumentalist compatibilists.

Gerald Gaus argues in *The Order of Public Reason* (2011) that the reactive attitudes played an integral role in the evolution of human cooperative society. This argument contains an important compatibilist insight, but Gaus does not put it to that end. His concern is with the justification of political coercion given a general liberal presumption in favor of individual freedom. Gaus's work is therefore more prominent in political philosophy than in compatibilist circles. The next paragraphs briefly summarize the relevant core of his argument.

Gaus argues that cooperative society cannot be developed or maintained based on instrumental rationality. His principal targets are contemporary Hobbesian political theorists such as David Gauthier (1986). The Hobbesian project is to instrumentally ground political morality: if individuals can rationally reason their way to accepting plausible principles of justice, then there is an elegant explanation for why (among other things) the state can be morally authoritative. The basic challenge for this project lies in puzzles of collective rationality, prominently the prisoner's dilemma. The main point is that there are many scenarios in which it is the instrumentally best for a self-interested agent to defect from a cooperative scheme. This is demonstrated by a *dominance* principle: in a simple two-person case, our agent can expect that their counterpart will either cooperate or themselves defect. If they cooperate, then our agent could take advantage of them by defecting. If they defect, then our agent would be better off defecting as well, rather than being taken advantage of.

Despite the individual rational dominance of defection, the overall results of cooperation are in many cases better. And obviously real people manage to cooperate all the time. So either people are not instrumentally rational, or else society can be arranged such that prisoner's dilemmas are transformed into situations in which cooperation is rational. The intuitive solution is a system of law and order that reliably punishes defectors: if the costs of defection are raised such that cooperation is rational, then

<sup>9</sup>Barrett (2020), Miller (2017), and McGeer (2014).

<sup>10</sup>I will focus mainly on practices of punishment, as the moral stakes of punishment are especially high. But various positive and negative reactive attitudes also play important roles in minimal morality. Thanks to a referee for encouraging clarification here.

all are better off. The difficulty is that any such scheme itself requires cooperation. There are always opportunities for corruption and free-riding in the enforcement of law and order, and this tends against stable cooperation. The dialectic at this point becomes very involved; suffice to say that Gaus argues (I think convincingly) that even very sophisticated Hobbesian theories ultimately fall prey to the same basic problem: there is no way to rationally overcome – that is, design a solution for – the basic dilemma for rational cooperation.

Gaus's key insight is that, although individual rationality may not favor cooperation, *groups* that somehow overcome this barrier will be significantly evolutionarily advantaged over those who do not. Evolution, so to speak, exhibits a broader rationality. The earliest form of cooperation is in kin groups. Indeed, the idea of narrowly individualistic human rationality is mostly figmentary, because the evolutionary ancestors of humans already developed kin loyalty networks.<sup>11</sup> Loyalty simply is the disposition to cooperate, and even sacrifice for, others in the social group. Loyalty can be modeled as arising like any other adaptively useful trait: there is a certain random spread of all sorts of different dispositions in a large population, and those that manage to form stable cooperative groups thrive (and thus reproduce) on the basis of strength in numbers.

As groups become larger – and larger groups tend to dominate smaller groups, so size is adaptive – cooperation requires moving beyond kin loyalty toward a proper *social morality*. Kin loyalty is compatible, among other things, with blood feuds that weaken the overall group. Social morality (rules, norms, traditions) enables group cohesion at larger scale. Stabilizing social norms, Gaus argues, requires that norms be internalized by “rule-following punishers” who are willing to enforce the rules even when this is not in their self-interest.<sup>12</sup> None of this is instrumentally self-interested in any narrow self-interested sense; evolution has simply hardwired the disposition to follow rules and to punish rule breakers.<sup>13</sup>

How are these pro-social dispositions internalized? For Gaus, the Strawsonian reactive attitudes play the perfect role: they are the moral-psychological profile of rule-following punishment. Moral emotions, on Gaus's view, are the motivational materials necessary for making demands even when it is not prudentially rational to do so: we simply are angry or resentful, and “emotions also typically have implications for appropriate action.” Put otherwise, emotions have (or are) action-tendencies (ibid: 189). On this picture, moral emotions are evolution's answer to the prisoner's dilemma – reactive attitudes yield (or perhaps constitute) reasons for enforcing social rules. This leads Gaus to remark that “When Strawson says that the ‘existence of the general framework of attitudes is something we are given with the fact of human society,’ he presents us with a deeper truth than even he realizes: human society would not even be possible without this framework” (ibid: 193).

I have not yet said much about the reactive attitudes themselves, nor about Strawson's own argument. I find it illuminating to approach the standard Strawsonian points from this angle. The reactive attitudes are a set of interpersonal responses to displays of good or ill will – gratitude, resentment, anger, praise, blame, and so on. Patterns of reactive attitudes are part of the “facts as we know them.”

<sup>11</sup>See Gaus (2021: 22–30) for related discussion.

<sup>12</sup>Gaus (2011: section 7).

<sup>13</sup>Of course, *given* these dispositions, cooperation is rational. But it is evolved dispositions that stabilize the system.

And our commitment to “ordinary interpersonal relationships” – those characterized by the reactive attitudes – is “too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction” might lead us to abandon our commitment to ordinary relationships (Strawson 1974: 12). (The general theoretical conviction being the incompatibility of determinism and moral responsibility.) The reactive attitudes are “part of the general framework of human life, not something that can come up for review” (ibid: 14).

This suggests a transcendental argument.<sup>14</sup> In short, the existence of human society is a fixed point. Human society as we know it is constituted by ordinary interpersonal relationships. Reactive attitudes are responses to good or ill will.<sup>15</sup> If actions could not be taken as displays of good or ill will because they only display preexisting causal chains, then we would thereby repudiate and abandon ordinary interpersonal relationships. But this, Strawson says, we cannot do. So the deterministic thesis that all actions have distant causal springs has no implications for actual social practices.

The obvious worry about this argument concerns its normative force. What do claims that are seemingly about human psychology have to do with the morality of blame and punishment? Even if it is true that we cannot abandon ordinary interpersonal relationships as Strawson construes them, what if they are morally inappropriate? Can interpersonal relationships be reformed such that the reactive attitudes, or at least the harmful ones, no longer have a central place?<sup>16</sup> Even if reform is impossible, are the reactive attitudes are a regrettable aspect of human nature?

The evolutionary argument derived from Gaus provides a powerful answer to this set of questions. The problem with the bare transcendental argument is that it seems to simply cut short moral justification of responsibility practices. The evolutionary argument makes the point that responsibility practices are given with the fact of human society, but adds the further claim that responsibility practices (characterized by reactive attitudes) are instrumentally *necessary* for the creation and maintenance of human cooperative society. Instead of denying that we are capable of seriously contemplating a theoretical (moral) challenge to responsibility practices, this raises the price of abandoning responsibility practices: if we were to abandon reactive attitudes, we would also abandon cooperative society as a whole. And this seems decisively morally bad. If so, then the necessary social and conceptual ingredients for cooperative society are instrumentally justified.

I will emphasize below that this argument only holds for the *minimal* necessary responsibility practices and is therefore compatible with radical critiques of actual responsibility practices, including critiques rooted in skepticism about free will. But it will help to develop the idea further by considering some nearby contrasts and objections.

I anticipate that the primary response to this proposal, as with other instrumentalist or two-level justifications of moral responsibility, will be that it invokes reasons that are “not even the right *sort* of basis” for justifying responsibility practices (Strawson 1974: 4). After all, the original (“optimist”) compatibilist approach that Strawson eschewed took off from the social efficacy of blame and punishment. Why should any instrumentalist account not be dismissed out of hand?

<sup>14</sup>See Coates (2017), Hieronymi (2020), and Russell (2021) for discussion. This is obviously not the only or even the dominant way to understand Strawson’s argument.

<sup>15</sup>Strawson’s famous argument for this is that when we *do* excuse or exempt individuals from moral responsibility, we do so because their actions did not in fact display ill will.

<sup>16</sup>See Pereboom (2001) for an influential development of this idea.

A small group of philosophers have recently argued that we should not be so quick to reject instrumentalist considerations as wrong kinds of reasons, at least not in Strawson's voice.<sup>17</sup> Their first point is to distance contemporary instrumentalist or consequentialist approaches from the views that Strawson likely had in mind. Barrett (2020) argues that Schlick (1939) and Smart (1961), two of Strawson's consequentialist "optimists," have unnecessarily blunt consequentialist arguments. Schlick argued as a *descriptive matter* that the point of our practices of punishment and blame is to spur moral reform. Strawson's characterization of the reactive attitudes is far more persuasive as an account of normal practice. Smart adopted a more plausible normative approach, arguing that we *should* punish with the intention of yielding good results, even if in practice people mostly punish in a backward-looking way. Barrett argues that this is not the right consequentialist conclusion: effective social regulation seems to empirically require backward-looking practices that maintain a norm of generally punishing violations of social rules.<sup>18</sup>

This can be pushed further. McGeer (2014) argues that Strawson should be read as anticipating "sophisticated consequentialism" of the kind advocated by Railton (1984) and Pettit (2012). Sophisticated consequentialists are similarly motivated by the idea that directly attempting to promote the best consequences will not work well in practice.<sup>19</sup> Their proposal is that we should adopt the *dispositions* that will have the best results. The best disposition will generally be based in social rules and norms that do not directly reference consequences. However, it may be worthwhile to refer back to forward-looking consequentialism in situations that are unusual or weighty. Sophisticated consequentialism recommends using two different deliberative standpoints: a backward-looking standpoint for everyday moral reasoning, and a forward-looking standpoint reserved for special circumstances. This is a two-level justification for moral responsibility (and everything else), but one in which the practical level is grounded in the theoretical level.

This has some traction on Strawson's appeal to reactive attitudes: the reactive attitudes are understood as a justified part of normal, routine moral deliberation and phenomenology, while the overall justification of moral responsibility practices would come at the more reflective, forward-looking level.<sup>20</sup> But there is a significant gap between this picture and Strawson's argument. Strawson argues that the general framework of ordinary interpersonal relationships is beyond questioning. Sophisticated consequentialism holds that we should not usually reason in a forward-looking way, but should support dropping into the consequentialist mode whenever there is time and opportunity for reflection. There are two perspectives, but only one normative framework: reflective deliberation is prior to habitual heuristics.

The evolutionary argument achieves a better theoretical fit by moving to the social level. Evolution is instrumentalist: the reason we are instilled with reactive attitudes is because they promote effective social cooperation. This fits with Strawson's claim – emphasized by consequentialist readers – that "It is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes." What is wrong, Strawson then says, is to forget that moral responsibility practices "do not

<sup>17</sup>Barrett (2020), Miller (2017), and McGeer (2014).

<sup>18</sup>See Barrett (2020: 6–9).

<sup>19</sup>The so-called "paradox of hedonism" is an instance of this general structure.

<sup>20</sup>This is notably akin to Strawson's distinction between ordinary reactive attitudes and the "objective stance" that we sometimes take up.

merely exploit our natures, they express them. Indeed the very understanding of the kind of efficacy these expressions of our attitudes have turns on our remembering this” (Strawson 1974: 27). Put differently: the ultimate purpose of the very framework of reactive attitudes may be instrumental, but our nature – importantly including the reactive attitudes of moral responsibility – is constituted within the evolutionary framework. This explains why Strawson finds something contradictory in consequentialist explanations of interpersonal relationships, the very purpose of which is to transcend instrumental rationality.

I think that he does not make the necessary distinction between justification *within* a practice and justification *of* a practice. The social evolution argument shows why there must be a non-instrumental framework of interpersonal morality, but does not rule out skeptical critique *of the framework*. Now, Strawson says that there is “endless room for modification” of responsibility practices “internal to the general structure” of reactive attitudes (1974: 25). So he does not mean to adopt a conservative view on which responsibility practices cannot be reformed, perhaps even radically reformed. But it is difficult to distinguish between internal and external justifications or critiques of moral responsibility practices. Strawson’s critical targets are instrumentalist optimists and incompatibilist pessimists. But we just saw that Strawson does not deny the relevance of the “efficacy” of responsibility practices to their justification. And some of the most influential presentations of incompatibilism begin with the familiar idea that involuntary actions are not blameworthy, and then attempt to generalize to the idea that all actions are involuntary (and thus not blameworthy) without libertarian free will.<sup>21</sup> What exact sort of criticism is acceptable?

It is notable in this context that Strawson generally talks about the impossibility of *completely* “abandoning” reactive attitudes, or their “total decay or repudiation” (1974: 14, 19). If the key point is that reactive attitudes cannot be abandoned wholesale, then the distinction between internal and external justification may be misleading. Similar arguments can be used to either revise or abandon moral responsibility practices. The social evolution argument suggests a principled explanation of why revisionism is acceptable but repudiation is not. In short, ordinary interpersonal relationships may be given as part of human nature and necessary for cooperative society, but we are not stuck with the particular social forms that we have now. We should be free to make any reforms for any reasons, so long as they do not violate the *minimal* conditions of cooperative society.<sup>22</sup> The evolutionary argument explains why reactive attitudes would arise, but evolution is a blunt tool, prone to overshooting its goals: we may have evolved to be far more punitive than minimal morality requires.

This suggests a progressive rather than incompatibilist deployment of deterministic argument. A critical interlocutor might accept that responsibility practices characterized by reactive attitudes play an ineliminable role in the development and maintenance of society. Unless the critic is also prepared to accept both political and social anarchism, this provides a sufficient reason to endorse some set of reactive attitudes – this line of justification, so far as it goes, is immune to determinism. But the critic should insist at this point that contemporary society is nowhere close to the minimally retributive

<sup>21</sup>Pereboom’s “four case argument” (2001) is a very explicit move from familiar (seemingly “internal”) intuitions to incompatibilism. Also see Watson’s (1987) famous discussion of the case of Robert Harris as an illustration of the power of causal history to soften reactive attitudes even in egregious cases.

<sup>22</sup>“Social Morality and Individual Ideal,” also reprinted in Strawson (1974), is suggestive here. Both Hieronymi (2020: 28) and Gaus (2011) are influenced by this essay.



morality. There may be many people who we could treat less harshly, and others who we could excuse from responsibility some or all of the time, without endangering social cooperation. (And similarly for positive reactive attitudes: one could surely have a cooperative society without the belief that anyone can deserve to be a billionaire.) At some frontier, abandoning further reactive attitudes would seriously harm ordinary interpersonal relationships and the fabric of cooperative society. Perhaps the best strategy would therefore be incremental: we should experiment with being more humane toward all, and quicker to treat various phenomena (mental health is a prominent case) as providing a good excuse or exemption from reactive attitudes.

It is worth noting how different this style of argument is from classical compatibilism. For consequentialists like Schlick or Smart, moral responsibility practices either reflected or were required by maximizing consequentialist logic. The argument here, while displaying an underlying instrumentalism, is not maximizing in nature. The idea of minimal morality is that there is a threshold of necessary backward-looking reactive attitudes that a stable society must. Of course it is hard to say where that threshold lies; hence the incremental progressive strategy. But minimally retributive morality is not *optimally* retributive morality: it is coherent and in fact seems overwhelmingly likely that truly minimal morality would leave some social welfare on the table, from the consequentialist perspective. The consequentialist would then be free to use this as a rationale to argue for more punishment (or more rewards for the high achieving, or whatever). The desert skeptic might leverage their belief that such punishment is (or should be) in some sense unfair or ill-fitting to argue against such a policy. Both positions are cogent: this sort of disagreement is intelligible even if both parties share a compatibilist view based on the necessity of *some* set of reactive attitudes for any society at all.

### 3. The structure of minimal compatibilism

The foregoing argument was situated largely at the second level of a two-level justification of moral responsibility practices. Its upshot is that, even accepting the incompatibilist premise that there is a general tension between determinism and moral responsibility, there is a strong rationale for maintaining a significant scope of moral responsibility practices, characterized by the reactive attitudes as non-instrumentally motivational backward-looking reasons to hold others accountable. This argument can be made while also holding demanding fundamental reforms of moral responsibility practices, including on desert skeptical grounds: if there is a *prima facie* conflict between determinism and backward-looking reactive attitudes that is only defeated by the minimal morality argument, then the desert skeptic is free to leverage their argument against all practices that go beyond truly minimal morality.

In this section I go further by arguing that this account is consonant with prominent compatibilist approaches – it can explain both the indispensability of reactive attitudes and the scope for their moral reform. I will focus especially on response-dependent accounts of moral responsibility: if a response-dependent account is subject to the kind of moral reform in the way I have in mind, I believe the other primary compatibilist contenders will clearly have the same space for moral reform.<sup>23</sup> David Shoemaker

<sup>23</sup>To mention how this would go: reason-responsiveness theories (e.g., Fischer and Ravizza 1998) could be reformed by increasing the demandingness of the relevant standard of responsiveness to reasons until only the minimal necessary set of reactive attitudes reasons; mesh theories (e.g., Frankfurt 1988, 1999) could be reformed by increasing the demandingness of the relevant agential mesh; fittingness-first accounts

(2017) has prominently advanced a response-dependent account argument, pressing an analogy to humor.<sup>24</sup> What is funny is determined by the norms of the practice. It does not make sense to say that a joke is not funny simply because finding it funny is morally bad in some respect (although it does at least make sense to say that one should not *laugh* if laughing is a morally bad act). And it may be that our comedic practices are so disparate and seemingly inconsistent – we find so many different kinds of things funny, but other superficially similar things not funny – that the only thing to say is that what *is* funny is determined by what we *find* funny, and no deeper theoretical analysis is possible.

However, it is perfectly cogent, and indeed familiar, to argue that the appropriateness standards of humor ought to be reformed for moral reasons. For instance, jokes at the expense of the mentally impaired have been significantly pushed out of socially acceptable humor in recent decades, because they came to be seen as cruel. Avoiding cruelty justified changes to comedic practice. Moral responsibility practices, in the same way, can be constituted by response-dependent norms of fittingness, while these norms can themselves be subject to justified moral reform.<sup>25</sup> One part of this, presumably the first part, is convincing people that they should not laugh or publicly display amusement at certain things. This is compatible with at least Shoemaker's version of response-dependence, which is about funniness rather than displays of amusement (2017: 16). But if this project is successful, such that people more broadly become disposed not to laugh at certain things, they will presumably eventually come to lose (or not form, in younger generations) the disposition to find the jokes in question funny, and perhaps eventually not see them as jokes at all.

This approach is especially congenial for response-dependent theories, because those views are otherwise especially prone to an objection that an incompatibilist is likely to make quite broadly: that compatibilists are simply *conservative*, willing to protect any (or at least too many) status quo practices. Shoemaker is at some pains to distance his view from a “dispositionalist” account, on which the funny or blameworthy is *simply* what we are disposed to laugh at or blame. This would entail that, if we began blaming young children or the severely cognitively impaired, it would simply *become* fitting to do so.<sup>26</sup> Instead, his response-dependent concepts are “thoroughly normative” (Shoemaker 2017: 4) – in particular, the blameworthy is whatever *merits* anger: the angerworthy (ibid: 508). What explains the scope of angerworthy things? Nothing more than that they are “the sorts of properties to which we humans are built to respond with a heated demand for acknowledgement or a tendency to retaliate. There is no better way to explain the motley collection of blameworthy fitmakers otherwise” (ibid: 510).

Peter Vallentyne (1996) poses a relevant dilemma for response-dependent accounts. A response-dependent account is “rigid” if appropriate responses are fully fixed in advance. In this case, the normative work is done by the constraints on responses rather

---

(Howard 2018; McHugh and Way 2016) could be reformed much in the same way as I go onto describe in connection with response-dependence.

<sup>24</sup>Shoemaker (2017).

<sup>25</sup>Of course, some people in many societies already reject punishment. But I assume that widespread (not universal) acceptance and participation in punishment practices (etc.) suffices to settle the appropriateness norms of moral responsibility. Thanks to a referee for pressing for clarification on this point.

<sup>26</sup>See Todd (2016) for a nice statement of this objection, pressed broadly against those who take up the “Strawsonian reversal” on which blaming practices are more basic than, and somehow fix, the appropriateness conditions for blame.

than the responses themselves; an “ideal observer” moral theory is response-dependent, but only superficially so if the ideal observer always chooses (by definition) on the basis of maximizing utility (or whatever). On the contrary, a more deeply response-dependent account that does not fix appropriate responses in advance leaves open the kind of contingency just mentioned: the account may approve of unsavory outcomes if that is the way the responses turn out. It is not entirely clear to me how Shoemaker would respond to this dilemma: the way “humans are built” does not seem like sufficient normative protection against intuitively bad outcomes, but he does not supply a further basis for the “thoroughly normative” response-dependent properties.

For my part, I am happy to agree with Shoemaker to this extent: our practices simply have their own appropriateness standards, which may not be explicable by any appeal beyond what we are “built” to do and feel. But I deny that this implies overall normative conservatism. Why should we accept that the way we are built to blame is the *best* way to be built?<sup>27</sup> Put differently, appropriate moral blame may be response-dependent. This does not rule out changing the scope of appropriate moral blame *by changing our responses*. And patterns of responses can be changed by social reformers in all kinds of ways, for reasons that are not themselves part of the norms of appropriateness.

The way in which moral reform operates on this picture is important to observe. Say that a moral reformer is moved by the incompatibilist intuition that there is something wrong about holding someone responsible if they lack free will. This idea, as Strawson said, is outside actual accountability practices: it is alien to ask in any particular case whether a wrongdoer should be excused because of the general thesis of determinism. This is like the baseball player arguing his strikeout with the claim that there should be *four* strikes allowed; or, closer to reality, by arguing that baseball should adopt an automated system of calling balls and strikes, under which that *last* call would never have been made.<sup>28</sup> In both cases, there is no answer to be given about what makes ill-treatment (being blamed, being called out) appropriate other than the fact that that is how the practice is. But it is perfectly cogent for the baseball player, after the game, to advocate for a change of rules. And his reason might be that the change of rules will make the game more fun to play or more fun to watch. If he can convince enough people, the rules may change. That means the practice will have changed. The new rules will not make any reference to how fun the game is, any more than the old rules did: they will simply list out (and thereby constitute) a *new* way it is.

The incompatibilist critique of moral responsibility interacts with evolution because evolution (plausibly) tells us that some set of backward-looking reactive attitudes is necessary for having a functional society, regardless of any other basis for moral desert. It does not make sense to advocate for a social reform that would make society impossible – no more than it would make sense to advocate that baseballs should be made out of a rare and unobtainable material.<sup>29</sup> If the critic refrains from demanding the total

<sup>27</sup>This phrasing is reminiscent of Vargas (2013). This is not coincidental, but I am more aligned with Vargas’s title (*Building Better Beings*) than with his substantive position as I understand it. Roughly, Vargas argues for revising our *concept* of free will to fit with our practices; I argue for revising our practices to fit our existing concept – to *make* what is true inside the practice fit better with what is morally justified – within the limits of the broadly instrumental argument.

<sup>28</sup>Compare Hart’s classic discussion of the “internal point of view” of the law (Hart 2012; Kaplan 2023), and Rawls (1955).

<sup>29</sup>It is intelligible to claim that we should simply stop having a society, like we might simply stop playing baseball. But we can simply reject that proposal out of hand.

abandonment of reactive attitudes, and demands only reforms that are plausibly compatible with functional society, there is no problem. But there is also no more principled incompatibilism. It is not that determinism makes blame inappropriate, but rather we should make blame inappropriate in certain domains by altering our blaming practices themselves. We should only undertake these reforms if we have sufficient reason to do so. Most obviously, we do not have sufficient reason to undertake reforms that would lead to social collapse – this is the minimal morality argument. (This leaves other questions open: perhaps we also lack sufficient reason to undertake reforms because they would merely lead to worse social outcomes overall, or even because implementing the reforms would simply be too emotionally taxing.)

A desert skeptic who accepts the need for minimal morality is therefore pressed toward a reformist stance that may be radical in the kind of reasons it brings to bear but incremental in the demandingness of those reasons. Importantly, both before and after a reform succeeds, determinism will have no place in responsibility practices themselves. Those are the rules of the game, and the rules do not make mention of the reasons one might reform them.

#### 4. The stability challenge

I have emphasized the importance of the two-level justification of moral responsibility practices. Given that the second (instrumentalist) level determines the scope of justified reforms to the first (appropriateness) level, one might question whether the distinction can bear the weight I have placed on it. If we “see through” appropriateness norms to their instrumentalist justification, the original worry about “optimistic” compatibilism may recur: when I am punished, this seems to be done in the name of the social good in spite of the fact that I do not really deserve punishment.

Minimal compatibilism requires appropriateness norms to be focal and instrumentalism to be constrained to the discourse of moral reform of the practice. The challenge is that the instrumentalist rationale may come to dominate the appropriateness norms: if only minimal moral responsibility is truly justified, why not evaluate whether each and every instance of accountability is indeed necessary? Going this route would be ultimately self-defeating, if truly backward-looking reactive attitudes are necessary for stable social norms. But this does not make it less worrisome: the lesson of the evolutionary account is that we cannot reason ourselves into successful social norms, no matter how much we might want to. Can backward-looking appropriateness norms be subjected to forward-looking moral reform without inviting their collapse? If not, minimal morality would require some other ingredient be added (or retained) in order to maintain stability: an obvious candidate is the widespread belief in the kind of libertarian free will that would ground moral desert.<sup>30</sup>

There are several reasons to think that the two-level structure is sustainable. First, consider again the analogy to humor, this time focusing on professional comedy. The ultimate justification for comedy as a practice, or for an individual comedian, may be instrumental. If there were no professional comics, the world would contain less joy, insight, and so on. A particular comic may just need the job. If the job or

<sup>30</sup>This argument might lead to “illusionism” on free will (Smilansky 2000, 2001, 2022). Illusionism holds that libertarian free will does not exist, but that the widespread belief therein is necessary to stability and other social and personal goods. If this were correct, then the moral reforms I advocate would be sharply limited to the extent that they risk undermining the general free will illusion.

practice were not justified, it would be strange to maintain it. But a comedian who could only see reasons to please this particular crowd or influence them with some piece of social commentary or simply take their money would probably not be very successful; the practice of comedy would surely suffer if it were made up solely of such instrumentalists. Good comedy is about being *funny*, and being funny does not always conduce to one's instrumental benefit.

Nor, as I suggested above, does moral reform seem to bother the core appropriateness norms of comedy. Some people (including some comics) seem to imagine a moralizing comedic dystopia, where every joke is evaluated first for its moral acceptability. Indeed, if the two-level justification collapsed fully, it would not even be intelligible to evaluate funniness except by reference to moral justification. But appropriateness norms seem more robust than this. Controversies about the acceptable limits of comedy do not seem to overwhelm the question of what is funny by and large; it is perfectly intelligible to hold that a particular comic is offensive, and even hold that this justifies moral sanction or "deplatforming," while still granting that they are funny. (In practice, it seems that mostly those who are offensive *and* unfunny suffer much from moral sanction.)

The sports analogy can be put to the same end: athletes and fans routinely distinguish between a rule being justified and being rightly adjudicated. (It is especially interesting when pro athletes speak openly about their ultimate purpose being to entertain. This does not seem to conflict with playing to win.) And analogies from all aspects of life could be multiplied. It may be thought that moral desert and moral responsibility are more fundamental ideas. But what is the evidence that moral responsibility practices are more fragile than anything else? Ordinary moral discourse is obviously not much concerned with libertarian free will or any other philosophical account of moral desert. Perhaps this reflects a common tacit assumption of free will. But major religious doctrines have obviously long wrangled with the conflict between divine omniscience and moral desert; it is not as if the ideas have not had time to settle in. It seems more likely that core moral responsibility practices are simply not that sensitive to justificatory questions except at the margins. And at the margins, aspects of desert skepticism are themselves part of moral discourse: addiction, mental health disorders, and even a bad upbringing are taken seriously as (partial) defeaters of reactive attitudes.<sup>31</sup>

Perhaps the point can be made most simply with an example. Gary Watson (1987) famously recounts the gruesome story of Robert Harris, who seems entirely callous even as he commits murder. He then recounts Harris's almost unbelievably traumatic childhood. Watson and many readers, myself included, find the fittingness of harsh punishment to be at least attenuated after understanding the fuller biography. In this moment of reflection, I feel that our moral responsibility practices depend in some measure on a false ideal of responsibility. I *also* feel, at the same moment, that punishment is fitting: how could it be right for someone to not face the consequences of their heinous actions? The former feeling obviously threatens the latter, but it seems to me quite plausible that it only does so indirectly. I could not truly abandon the feeling of fitting punishment without immersing myself in an entirely different kind of social practice. And then it is apt to ask whether such immersion, and such a social practice, would be justified. If it is not, the only thing to do is to accept an abiding tension between an incompatibilist element in reformist thinking and the actual compatibilist reality of justified moral responsibility practices.

<sup>31</sup>See Shoemaker (2015) and Pickard (2017).

I cannot claim that such reflections are entirely decisive. The stability challenge is ultimately a sociological one. So perhaps it is best to present the final conclusion in a conditional form. *If* at least some backward-looking reactive attitudes, paradigmatically blame but including many positive and negative reactions, are necessary for a stable society, *and* this set of reactive attitudes can be morally reformed to progressively seek its minimally harmful core, then the account of minimal compatibilism that I have offered has several attractions. It retains the Strawsonian focus on the centrality of our “given” practices of reactive attitudes: it even retains the controversial Strawsonian claim that practices of praise and blame fix the standards of appropriate praise and blame. However, by emphasizing second-level moral justification and reform of social practices, minimal compatibilism avoids the implausibly conservative aspects of this view. What *is* appropriate does not settle what *should be* appropriate. I accept Strawson’s claim, folding in Gaus’s evolutionary argument, that the existence and basic character of our moral responsibility practices is not seriously up for question. But this can be combined, as I have tried to show, with the revisionary moral impulse usually associated with incompatibilists. It is only the minimal moral responsibility practice that cannot be questioned. Our moral duty is to seek that minimum.

**Acknowledgments.** I would like to thank Andrew Lichter, Luke Golemon, Anna-Bella Sicilia, and Max Kramer for their comments and discussion. I am especially grateful to Michael McKenna for his help throughout the development of this paper.

## References

- Barrett, Jacob.** 2020. Optimism about Moral Responsibility, *Philosophers’ Imprint*, 20.33: 1–17.
- Caruso, Gregg D.** 2021. *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice, Law and the Cognitive Sciences* (Cambridge: Cambridge University Press).
- Coates, D. Justin.** 2017. Strawson’s Modest Transcendental Argument, *British Journal for the History of Philosophy*, 25.4: 799–822 <<https://doi.org/10.1080/09608788.2017.1284647>>.
- Fischer, John Martin, and Mark Ravizza.** 1998. *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press).
- Frankfurt, Harry G.** 1988. *The Importance of What We Care about: Philosophical Essays* (Cambridge: Cambridge University Press).
- Frankfurt, Harry G.** 1999. *Necessity, Volition, and Love* (Cambridge: Cambridge University Press).
- Gaus, Gerald.** 2011. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, 1st paperback edition (Cambridge, MA: Cambridge University Press).
- Gaus, Gerald.** 2021. *The Open Society and its Complexities*, 1st edn (Oxford: Oxford University Press). <<https://doi.org/10.1093/oso/9780190648978.001.0001>>
- Gauthier, David P.** 1986. *Morals by Agreement* (Oxford [Oxfordshire]: New York: Clarendon Press; Oxford University Press).
- Harris, Sam.** 2012. *Free Will*, 1st Free Press trade pbk. edn (New York: Free Press).
- Hart, H. L. A.** 2012. *The Concept of Law, Clarendon Law Series*, 3rd edn (Oxford, UK: Oxford University Press).
- Hieronymi, Pamela.** 2020. *Freedom, Resentment, and the Metaphysics of Morals*. (Princeton: Princeton University Press).
- Howard, Christopher.** 2018. Fittingness, *Philosophy Compass*, 13.11: e12542.
- Kaplan, Jeffrey.** 2023. The Internal Point of View, *Law and Philosophy*, 42.3: 211–36 <<https://doi.org/10.1007/s10982-022-09461-x>>.
- McGeer, Victoria.** 2014. P. F. Strawson’s Consequentialism, in *Oxford Studies in Agency and Responsibility*, Volume 2 eds. David Shoemaker and Neal Tognazzini (Oxford: Oxford University Press), pp. 64–92 <<https://doi.org/10.1093/acprof:oso/9780198722120.003.0005>>.

- McHugh, Conor, and Jonathan Way. 2016. Fittingness First, *Ethics*, 126.3: 575–606 <<https://doi.org/10.1086/684712>>.
- McKenna, Michael. 2020. Punishment and the Value of Deserved Suffering, *Public Affairs Quarterly*, 34.2: 97–123.
- McKenna, Michael. 2021. Wimpy Retributivism and the Promise of Moral Influence Theorists, *The Monist*, 104.4: 510–25 <<https://doi.org/10.1093/monist/onab016>>.
- Miller, Dale E. 2017. “Freedom and Resentment” and Consequentialism: Why “Strawson’s Point” is Not Strawson’s Point, *Journal of Ethics and Social Philosophy*, 8.2: 1–23 <<https://doi.org/10.26556/jesp.v8i2.79>>.
- Pereboom, Derk. 2001. *Living without Free Will* (Cambridge: Cambridge University Press).
- Pereboom, Derk. 2021a. Undivided Forward-Looking Moral Responsibility, *The Monist*, 104.4: 484–97.
- Pereboom, Derk. 2021b. *Wrongdoing and the Moral Emotions*, 1st edn (Oxford: Oxford University Press) <<https://doi.org/10.1093/oso/9780192846006.001.0001>>.
- Pettit, Phillip. 2012. The Inescapability of Consequentialism, in *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams*, edited by Ulrike Heuer and Gerald Lang (New York: Oxford University Press).
- Pickard, Hanna. 2017. Responsibility without Blame for Addiction, *Neuroethics*, 10.1: 169–80.
- Railton, Peter. 1984. Alienation, Consequentialism, and the Demands of Morality, *Philosophy and Public Affairs*, 13.2: 134–71.
- Rawls, John. 1955. Two Concepts of Rules, *The Philosophical Review*, 64.1: 3 <<https://doi.org/10.2307/2182230>>.
- Russell, Paul. 2021. Responsibility Skepticism and Strawson’s Naturalism, *Ethics*, 131.4: 754–76 <<https://doi.org/10.1086/713954>>.
- Sapolsky, Robert M. 2023. *Determined: A Science of Life without Free Will* (New York: Penguin Press).
- Schlick, Moritz. 1939. *Problems of Ethics*, trans. by David Rynin (New York: Prentice-Hall).
- Shoemaker, David W. 2015. *Responsibility from the Margins* (Oxford: Oxford University Press).
- Shoemaker, David. 2017. Response-Dependent Responsibility; or, a Funny Thing Happened on the Way to Blame, *Philosophical Review*, 126.4: 481–527 <<https://doi.org/10.1215/00318108-4173422>>.
- Smart, J. J. C. 1961. Free Will, Praise, and Blame, *Mind* 70: 291–306.
- Smart, J. J. C., and Bernard Williams. 1973. *Utilitarianism: For and Against*. (Cambridge: Cambridge University Press).
- Smilansky, Saul. 2000. *Free Will and Illusion, Repr* (Oxford: Clarendon Press).
- Smilansky, Saul. 2001. Free Will: From Nature to Illusion, *Proceedings of the Aristotelian Society (Hardback)*, 101.1: 71–95 <<https://doi.org/10.1111/j.0066-7372.2003.00022.x>>.
- Smilansky, Saul. 2022. Illusionism, in *The Oxford Handbook of Moral Responsibility*, 1st edn, eds. Dana Kay Nelkin and Derk Pereboom (Oxford: Oxford University Press), pp. 203–C10.P129 <<https://doi.org/10.1093/oxfordhb/9780190679309.013.9>>.
- Strawson, Peter Frederick. 1974. *Freedom and Resentment and Other Essays* (Oxfordshire: Routledge), pp. 185–88.
- Todd, Patrick. 2016. Strawson, Moral Responsibility, and the “Order of Explanation”: An Intervention, *Ethics*, 127.1: 208–40 <<https://doi.org/10.1086/687336>>.
- Vallentyne, Peter. 1996. Response-Dependence, Rigidification and Objectivity, *Erkenntnis* 44.1: 101–12 <<https://doi.org/10.1007/BF00172855>>.
- Vargas, Manuel. 2013. *Building Better Beings: A Theory of Moral Responsibility*, 1st edn (Oxford: Oxford University Press).
- Vargas, Manuel. 2015. Desert, Responsibility, and Justification: A Reply to Doris, McGeer, and Robinson, *Philosophical Studies*, 172.10: 2659–78 <<https://doi.org/10.1007/s11098-015-0480-7>>.
- Waller, Bruce N. 2011. *Against Moral Responsibility* (Cambridge, MA: MIT Press).
- Watson, Gary. 1987. Responsibility and the Limits of Evil: Variations on a Strawsonian Theme, in *Perspectives on Moral Responsibility*, eds. John Martin Fischer and Mark Ravizza (Ithaca: Cornell University Press), pp. 119–48.
- Wegner, Daniel M. 2002. *The Illusion of Conscious Will* (Cambridge, MA: MIT Press).

Cite this article: Quigley, Travis. 2024. Compatibilism and Truly Minimal Morality. *Utilitas* 36, 323–337. <https://doi.org/10.1017/S0953820824000177>