In *Essays on Free Will and Moral Responsibility*. Edited by Nick Trakakis and Daniel Cohen. Cambridge Scholars Publishing (in press).

MORAL INFLUENCE, MORAL RESPONSIBILITY

MANUEL VARGAS

The traditional consequentialist model of responsibility holds that praise and blame are forward-looking attempts to influence agents in socially desirable ways. On this account, praise and blame derive their justification from their efficacy at facilitating desirable outcomes. In the literature this picture of moral responsibility—one I will call 'moral influence'—has been widely rejected as a failure. As P.F. Strawson famously put it, "[moral influence] is not a sufficient basis, it is not even the right *sort* of basis, for these practices as we understand them".[1] The consensus—and it is virtually unanimous among philosophers of free will and moral responsibility—is that moral influence theories have little to offer in the way of an adequate theory of moral responsibility.

In this paper, I aim to identify an important insight that rests at the core of traditional moral influence theories, and to develop that insight in a way that sidesteps the traditional objections directed against these accounts. The insight I aim to make use of is roughly this: the *justification* of our praising and blaming practices derive, at least in part, from their effects on creatures like us. The appeal of this justificatory strategy is that, if it works, it provides a way to justify our responsibility-characteristic practices in a way not dependent on traditional debates about the metaphysics of free will and responsible agency. Indeed, in providing an independent justification for our responsibility-characteristic practices, the account undercuts some of the motivation for skepticism about moral responsibility. So, my aim is not to condemn the moral influence approach but to praise its spirit.

This paper has five parts. In the first part I canvass the main objections to the moral influence theory. In the second part, I develop an account that rescues what I take to be the kernel of truth in moral influence theories. In the third part I describe some of the burdens and limits of the account I offer in part two. In the fourth part I show how my modest redeployment of the moral influence idea is immune to the traditional objections. In the fifth and concluding section, I consider a final objection about the implications of this sort of account.

1. Some Traditional Objections to the Moral Influence Approach

> The difficulties with this theory are, I think, well known.
> —T.M. Scanlon, "The Significance of Choice"[2]

According to the moral influence account of responsibility, the attitudes and practices characteristic of moral responsibility are justified because such practices influence or pressure agents to behave in morally desirable ways.[3] One feature of the theory is that it seems to draw the distinction between responsible and non-responsible agents in approximately the right place, and it does so without appealing to what P.F. Strawson later described as "panicky metaphysics". So, for example, in cases where there is no point to influence or pressure—e.g., when an agent is coerced, has insufficient knowledge, or fails to be sane enough—a moral influence account maintains that the considered agent ought not be held responsible. Where influence can be effective, we should hold people responsible.

Despite this initial sheen of plausibility, the disrepute into which moral influence accounts have fallen requires that any rehabilitation of it must directly address the traditional objections. So, I'll begin by presenting the objections. Later, I will reply to them on behalf of the account I offer. What I will not do is to take a stand on is whether the standard criticisms of the traditional view *ought* to have felled traditional moral influence accounts; happily, our present purposes constrain the degree to which we must look backwards.

* * *

Over the past forty years, numerous criticisms have been directed against the moral influence view of responsibility.[4] One of the more serious objections to moral influence (MI) theories is this: they are too coarse-grained to make the distinctions required of a theory of moral responsibility, despite the initial appearance of plausibility to which I recently referred. This criticism is twofold: MI theories cannot make suitable distinctions among kinds of agents, and relatedly, moral influence itself cannot be distinguished from other kinds of influence.

Take the matter of distinguishing between intuitively responsible and intuitively non-responsible agents. MI theories maintain that agents are responsible when they can be influenced in the ways characteristic of holding people responsible. On this account, the threat or promise of our anger, indignation, gratitude, praise, blame, punishment, or reward is the substance of moral responsibility. But, if it is mere susceptibility to influence that marks out responsible agents from non-responsible agents, then we do not have any principled way to distinguish intuitively responsible agents (normal adults, for example) from intuitively non-responsible agents (infants, most non-human animals). This is because members of both groups can be moved by a range of 'influencing' behaviour (cajoling, threatening, enticing, and so on). So, even if an MI account can explain the pointlessness of blaming someone who acted out of ignorance, it cannot explain why we should, in the ordinary case of blaming, hold normal human adults to a kind of moral standard we do not intuitively think appropriate to hold of infants and most (or perhaps all) non-human animals. It thus fails to adequately account for the difference between responsible and non-responsible agents.

Now consider the other aspect of the coarse-grainedness objection. This complaint holds that MI accounts do badly in distinguishing between moral and non-moral forms of influence. If holding someone morally responsible just is to treat them in ways that would influence them to behave better, then we have no way to distinguish intuitively genuine blaming from feigned blaming. Indeed, we would have no way to distinguish moral influence from a range of manipulative behaviour that aims to shape others and their actions without any attendant moral judgement. The natural place to look for a distinction between moral and non-moral influence is internal to the act of influence (say, an instance of moral blaming). But on the moral influence account it is difficult to see how there could be a relevant "internal" feature, given that MI accounts construe responsibility and its attendant judgements and practices in terms of some external relation (namely, whether the influence is efficacious). So, here too a moral influence theory is simply too ham-fisted. Its failure to distinguish between kinds of agents and kinds of influence shows it to be an inadequate theory of responsibility.[5]

A second major objection holds that MI conflates being responsible with judgements about the appropriateness of holding responsible. On a standard moral influence theory, an agent's *being* responsible is fixed by facts about when and/or whether it is appropriate to hold the agent responsible—that is, whether we can influence an agent (or others) to behave in a suitable way. However, at least in commonsense moral thinking, whether someone is responsible and whether it is appropriate to hold that person responsible come apart. Suppose that we have a policy of never holding people responsible for, say, the first impolitic remark they make. Now suppose that we make an arbitrary exception, and hold one and only one person responsible the first time he or she makes an impolitic remark. It looks as though we can say of this case that even though the agent is, in fact, morally responsible for the first impolitic remark, it would be inappropriate to hold him or her responsible. Thus, a theory that collapses the issue of when someone is morally responsible with the issue of when it is appropriate to hold someone morally responsible will be unsatisfactory.[6]

A third objection is that MI accounts fail to accurately describe *how* we hold people responsible. It is entirely compatible with a moral influence account that one never need experience what Strawson called *the reactive attitudes*—the responsibility-characteristic attitudes of resentment, indignation, gratitude, and so on—and could instead feign these things as part of an attempt to influence others. On the moral influence account, genuine resentment, indignation and so on, are never actually required. In fact, a MI theory might recommend (or perhaps even require) something like an emotionally disconnected, almost therapeutic approach to influencing others by the most expedient means. However, reflection on how we in fact hold people responsible shows that "blame-related responses all involve something like hostility towards the subject; whereas a [moral influence] therapist, though he may have to feign ill-feeling for therapeutic purposes, can in fact be in a perfectly sunlit frame of mind."[7] Even if the practice of holding people responsible *sometimes* amounts to an attempt to influence people, it is surely a mistake to claim that we are *always* attempting to influence others.

A fourth objection is that MI theories mistakenly identify the moment of moral concern, always locating it in the present or the future, and never in the past. Sometimes, however, assignments of responsibility are backward-looking. That is, they are assessments of the way an agent was, and have little or nothing to do with the influence that the reactive attitudes or associated practices might have on this agent or others in the future. Consider gratitude on a moral influence account. I cannot just be thankful for something you have done in the past. For my gratitude to be justified, it has to be the case that my expression of gratitude would encourage you in the right way. This is, by itself, deeply puzzling, but it also suggests a wider problem with cases where someone is beyond the sphere of influence. Surely we can have responsibility-characteristic attitudes such as gratitude toward parents, grandparents, friends, and mentors even if they are dead or otherwise incapacitated. Even if it turned out that such attitudes had *some* justification because of their effects on the living, this again seems to be the wrong kind of justification for gratitude. Because moral influence accounts are exclusively present or forward-looking, they fail to capture our often legitimate concern for the past.[8]

A fifth and final objection is a simple but not insignificant one. It is the objection that moral influence theories are unacceptably tied to consequentialism. If one finds consequentialism troubling on independent grounds, any theory of responsibility that presupposes some form of consequentialism will seem troubling for that reason. A somewhat more nuanced version of this worry is worth mentioning here as well. Given the contentious nature of normative ethics, a theorist of responsibility should treat it as a desideratum that any proposed account of moral responsibility be somewhat insulated from commitments to a specific theory of normative ethics. Given this desideratum, moral influence theories are problematic not because of consequentialism as such, but because they imply a commitment to a specific moral theory.

Individually, and sometimes jointly, these five (or so) objections have been widely taken to show the inadequacy of the MI account as a theory of moral responsibility.

2. The Requisite Brief Aside on Methodology

> …the idea of Free Will seems involved in a particular way in the moral ideas of Common Sense.
> —Henry Sidgwick, *The Methods of Ethics*[9]

In what follows, I propose an account of the justification of our responsibility-characteristic practices of moral praise and blame that makes use of some ideas traditionally associated with moral influence accounts. My interest here is *not* folk-descriptive, or what I have elsewhere called a *diagnostic* project.[10] That is, I am not interested in an account strictly beholden to the limits of what we happen to believe about moral responsibility. Rather, I am interested in an account that specifies what we *ought* to think about moral responsibility, at least in our capacity as theorists.

My focus on a prescriptive account, one that is prepared to revise ordinary intuitions, can sometimes raise the worry that the theory is immune to refutation. The worry is that any objection will be dismissed on the grounds that it presumes a non-revisionist theory of responsibility.[11] In reply, note that no account escapes a minimal standard of non-contradiction. So, it is not as though the account is completely immune to the possibility of critique. The account is vulnerable to more substantive critique, however. For example, one would have reason to reject it if it was committed to something that we had good independent reason to think was false. For example, if the picture of agency presumed by the theory were plainly false in light of, say, research in social psychology, this would be a compelling objection to the theory. Third, and more generally, it would be a *prima facie* problem for the theory if it had counterintuitive results that lacked a principled explanation. So, if the theory maintains that the behaviour of chinchillas could be blameworthy, but this upshot had no principled explanation and no account of why we should abandon the ordinary intuition, this would be an objectionable result.

In sum, departures from commonsense are not troubling if either of two conditions hold for the considered case: (1) commonsense relies on a metaphysically implausible notion of agency, or (2) an alternative account offers a comparatively better justification (as opposed to no justification or an implausible justification) for revision than a non-revisionist account can offer. For these reasons, openness to revisionism does not, by itself, guarantee that the theory is immune to refutation.

3. Outlines of a Theory of Moral Responsibility

In providing an account of an important aspect of moral responsibility, I wish to be clear about the account's scope and aspirations. Any complete theory of moral responsibility—something I will call *A Grand Theory of Responsibility*—will require the integration of at least three kinds of subordinate or constitutive theories. These theories are:

> (1) *A theory of responsible agency*, or an account of what sorts of agents the justified norms of responsibility apply to. It is doubtful that rocks, agave plants, or chinchillas are the correct targets of norms governing moral praise and blame. A theory of responsible agency provides a general account of how to distinguish between right and wrong targets for genuine, desert-imputing moral blame and praise, which in turn allows us to distinguish between various limit cases (children, psychopaths, paranoid schizophrenics, and so on).
>
> (2) *A theory of the responsibility norms,* or an account of the content of the justified responsibility norms. Where a theory of responsible agency tells us who is an appropriate candidate for ascriptions of responsibility, a theory of the responsibility norms provides an account of the norms that govern the application of praise and blame among candidate agents. So, for example, that

> one is the right sort of creature to be subject to praise and blame does not settle whether or not one deserves praise and blame for some particular action. To determine whether praise or blame is deserved we need to appeal to some account of when responsible agents are to be praised and blamed, when an agent is excused for some action and when an agent has been negligent. This is the work for a theory of the responsibility norms.
>
> (3) *A theory of the justification of the responsibility norms*, or an account of the basis on which the responsibility norms are justified. The norms of responsibility are entitled to bind us only if they are justified. What an account of the justification of the responsibility norms provides is an explanation of the normative basis of responsibility norms.

However, in providing an account of the justification of the responsibility norms it is likely that we will also appeal to a fourth sort of account:

> (4) *A theory of the aim of the responsibility system*, or an account of what the system of norms, practices, attitudes, judgments, and concepts associated with responsibility is properly understood as directed at. It might turn out that there is nothing that substantively unifies or organizes *the responsibility system*, that is, our justified responsibility-characteristic practices, attitudes, judgments, concepts, and norms. However, if there is some such organizing aim, principle, target, or role for the responsibility system, then an account of it is likely to be relevant to the account of the justification of the responsibility norms. An account of the aim of the responsibility system might also provide for a kind interrelation and systematicity between the various aspects of a Grand Theory of Responsibility.[12]

Conceiving of theorizing about moral responsibility along these lines helps make clear the explanatory burdens of any partial account of moral responsibility (by which I mean anything short of a Grand Theory of Responsibility). The explanatory burdens of each subordinate theory are important to keep in mind, as I am not attempting to offer a Grand Theory of Responsibility. I will say only a little about the nature of responsible agency and the content of the responsibility norms (1 & 2, above). This is because the idea of moral influence is badly applied when conceived of as a theory of those things. The kernel of truth in MI accounts is not so global in its reach. *All MI properly provides is an account of the justification of the responsibility norms* (3, above). This is important work for a theory of moral responsibility, and it can be done well by a properly constrained conception of moral influence. However, it is still only a partial account of responsibility, and one that is constrained in its deployment of a notion of MI. So, for example, when I appeal to some account of responsible agency or the content of the responsibility norms, I can do so without those accounts also relying on an idea of moral influence.

It is perhaps the most important flaw of traditional MI accounts that they attempt to generate a Grand Theory of Responsibility from the comparatively limited idea of moral influence. But a more modest scope of application for the idea is not enough. What is also needed is a refinement of the idea of moral influence itself. Recall my earlier characterization of traditional MI accounts as being committed to two related ideas: (i) praise and blame are forward-looking attempts to influence behaviour in socially desirable ways, and (ii) the justification of praise and blame comes from their effects. It is the second claim that I think is roughly right, though perhaps not in the way moral influence theorists have traditionally argued. There is, I think, a temptation on the part of moral influence theorists and their critics to think of the justification of moral influence in terms of the efficacy of particular tokenings of praise and blame, such as when, say, Lori criticizes Dan for being overly self-conscious or when we praise Michael for his copious feedback on a paper. A more compelling alternative is one that construes the justification for moral praise and blame as arising not at the level of particular interpersonal interactions but instead at the level of a general practice. In particular, the justification arises from the group-level effects of justified norms that are ubiquitously internalized by members of the community and regularly put into practice. This transformation of the idea of moral influence —both a scaling back of ambition and elevation of the source of justification from tokens to the in-practice effects of the system of norms of praise and blame as a whole— has several important consequences, not the least of which is that many of the traditional objections to moral influence accounts are dissolved.

To see how we might make good on this improved notion of moral influence, it helps to invoke an account of the aims of responsibility, as well as some account of responsible agency. On the view I favour, the aim of the responsibility system is to foster a distinctive form of agency in us, a kind of agency sensitive to and governed by moral considerations. Although our responsibility-practices and judgments may appear to function as a kind of coercive enforcement mechanism on behalf of morality, to think of responsibility in this fashion is to conflate effect and aim.[13] Instead, we should think of our responsibility practices, attitudes, and judgements as organized around the development and promotion of our moral considerations-mongering agency, and in particular the refinement of our sensitivity to moral considerations and the expansion of the contexts in which we reliably detect and act upon the relevant moral considerations.

This picture of the aim of the responsibility system makes a particular account of responsible agency seem appealing. That picture is one where responsible agency requires the presence and normal operation of various basic psychological features, including beliefs, pro-attitudes and intentions, but also responsiveness to moral considerations.[14] Thus, the distinctive mark of agents appropriately subject to the norms of responsibility is, roughly, the capacity to detect and regulate behaviour in light of moral considerations, where these capacities are indexed to facts about the agent's circumstances and the practical and justified interests that govern our ascription of capacities. I have argued for such an account (with various epicycles) elsewhere. Here, I will simply assume the adequacy of that account.[15]

Given a picture where we assume (1) that the *aim* of the responsibility system is to foster a kind of moral considerations-responsive agency, (2) that the correct account of *responsible agency* is one that broadly favours moral considerations-responsive agency, and (3) that a moral influence account is the correct account of the *justification* of the norms of praise and blame, then one might think justificatory role of moral influence should feed back into our account of responsible agency in a particular way. That is, one might think that a further condition holds on responsible agents: responsible agents must be influenceable. But, of course, agents responsive to moral considerations *are* influenceable precisely in virtue of their sensitivity to moral reasons. Indeed, it is the presence of this sensitivity that normally makes otherwise mere agents into responsible agents.[16]

So, on the picture we are leading up to—an initial account of moral influence as a theory of the justification of the responsibility norms—what we get is the following account of the justification of the norms of moral responsibility: we are justified deploying responsibility-characteristic practices where such practices, as a whole and over time, aid responsible agents to act in ways governed by moral considerations. Appropriately holding an agent responsible involves rightly regarding them as a responsible agent and correctly applying the justified norms of praise and blame, norms that derive their justification from their collective effects on fostering responsible agency.

To determine whether this provisional proposal of a more limited conception of moral influence might work, we need to know whether it is plausible to think that responsibility-characteristic practices might influence in the appropriate ways. The traditional model of moral influence gets this much right: typically, responsibility-characteristic practices—such as praising, blaming, punishing, and rewarding—work by providing external motivation for agents to track moral considerations and regulate their behaviour in light of them. Effective practices will exploit our psychology, and are largely parasitic on it. Typically, for creatures like us, praise encourages and blame discourages.

One consequence of regularly enforced norms is that individual agents typically come to internalize those norms. Ordinarily, there are clear deliberative benefits to having compliance with those norms become second nature, at least in contexts that regularly enforce then. When the norms are internalized, the agent need not deliberate about what to do, from the perspective of the norms. In the case of the responsibility norms, internalizing them helps to make assessments and choices in a way that permits the agent to reliably avoid sanction and reliably earn praise and reward. The upshot of this process, when it involves *justified* norms, is important. The result is an agent oriented towards tracking and responding to moral considerations. Moreover, once internalized and habitual, the threat of actual praise and blame need not play any active role in deliberation. The norms will oftentimes have a kind of motivational inertia in how they structure the perception of available courses of action, persisting even in the absence of external praise and blame.[17]

Moreover, internalization of the responsibility norms does more than structure the conception of action possibilities: it also structures the agent's self-assessments of praiseworthiness, blameworthiness, and desert. In doing this, internalized responsibility norms come to shape both the agent's prospective and retrospective self-assessments, which in turn license a range of evaluative and emotional responses from blame and recrimination to positive self-regard and self-satisfaction.

However, that responsibility norms can be internalized does not, by itself, suffice to ensure infallible compliance with the justified norms. Nothing in this account denies the possibility of akrasia with respect to internalized norms, or the possibility that the agent will decide that there is some reason that trumps the norms of responsibility. Moreover, it is always possible that the agent will imperfectly internalize norms, or that the norms an agent has internalized are not justified. My point here, though, is that practices of moral praise and blame can come to structure the deliberations of agents even when actual expressions of praise and blame are unlikely or absent.

This initial account of a more limited moral influence approach to moral responsibility requires augmentation. First, there is no reason to suppose that the influencing effects of praise and blame must be direct. That is, some apparently non- or even wrongly-influencing instances of responsibility-characteristic attitudes and practices may indirectly contribute to the general efficacy of the practices over time. And crucially, what justifies the responsibility system are its global effects—the concern here is for the norms governing our responsibility-characteristic practices and attitudes as a whole (I will return to this point in a bit).

To see why indirect effects might contribute to the justification of our responsibility practices, consider first a non-moral case. Let us suppose that the aim of the practice of football includes fun for the competitors and entertainment for the spectators. The rules of football may sometimes require games where a sequence of foul calls is neither fun for the competitors nor conducive to the entertainment of the fans. However, having a regular, stable system of foul calls in place surely contributes to the fun and entertainment of the sport over time. Analogously, there may be instances where my gratitude may fail to influence anyone in the proper fashion. Nonetheless, my gratitude can have an appropriate role, internal to the system of moral influence, because the prevalence of such attitudes and corresponding practices contributes to the efficacy and stability of the responsibility system over time.[18]

A second point to recognize is that our psychology puts limits on the justified norms of influence. As we have seen, the efficacy of the norms depends partly on their internalization. This internalization, however, relies on marshalling complex psychological forces. Since our psychologies are messy—that is, many of the mechanisms involved have functions and histories that do not neatly map onto the social roles we expect of people—it is likely that some of the psychological mechanisms on which responsibility practices and attitudes depend play diverse roles in our psychological economy. Anger, resentment, satisfaction, and so on can have both moral and non-moral roles in that economy. The various roles these attitudes play and the psychological mechanisms they rely on may impose limits or create psychological phenomena that—at least from a specifically moral perspective (or even a more limited perspective exclusively concerned with moral responsibility)—are undesirable or in tension with the distinctly moral roles for which those mechanisms have been appropriated. It would be a mistake to assume that the responsibility-characteristic attitudes exist solely as the substrate for our responsibility practices. Once we recognize this point, we have to allow for some slack between the mechanisms of influence and the efficacy of the outcome. We cannot suppose that the psychological mechanisms we rely on in influencing and being influenced are optimally efficient at motivation. Our complete psychological economy is too complex for that; the individual mechanisms of influence and motivation play various roles that are not likely to be optimized exclusively for the subtle and complicated mechanisms of moral appraisal.

A third way in which our moral psychology is relevant has to do with the possibility that some of the responsibility-characteristic attitudes may be unavoidable and largely unchangeable. For example, we may well discover that feeling resentment at being a target of apparently unjustified ill will is a largely implastic piece of our cognitive and affective architecture. If so, a theory of responsibility will have to allow for resentment, even if resentment generally fails to contribute to a system that fosters moral consideration-sensitive agency. There is no need to deny this possibility, for every theory of responsibility will be constrained by the limitations of human psychology. As even the most effective set of practices will contain concessions to our psychologies, the best we can hope for are practices that, *as a whole*, work reasonably well with and for creatures with psychologies like ours.[19]

This last point is an important one, and it returns us to a point I made in passing a few paragraphs back: the justification of our responsibility norms should be understood as the justification of a *network* of norms that underpin a web of practices, attitudes, and judgments. Individual practices or attitudes may not serve to influence a particular agent in a suitable fashion. However, if those practices or attitudes are necessary upshots of a psychology-dependent system that enables us to promote the relevant justified ends, then norms that respect that fact are perfectly acceptable and, indeed, required. What matters is the overall efficacy of the responsibility system in influencing us, and not a particular instance of holding someone responsible. The norms that are justified just are those norms whose currency in the psychology and practices of a community would, in fact, foster among us the kind of moral considerations-mongering agency that is the responsibility system's concern.[20] Plausibly, many of these norms will be expressed in practices that, in their exercise, conception and application have no element of immediate concern with influence.

This point allows us to see something about what I'll call *necessary inefficacies*, as distinct from indirect effects. What is distinctive about necessary inefficacies is that they are side-effects of an otherwise effective system. Whether something is a necessary efficacies or an indirect effect depends on whether it contributes to the aim of the responsibility system or whether it is a by-product of something else that is necessary for pursuit of the aim of the responsibility system. So, suppose we learn that moral revulsion can be jettisoned, psychologically and socially speaking. Further, supposed we learn that it does not directly contribute to the fostering of moral considerations-responsive agency. Before we recommend excision of it from our moral practices we would need to know if it indirectly contributes to the overall efficacy and stability of the responsibility system over time. If it does, then it may have a place in a justified system of practices in light of its indirect effects. If it does not, then we still need to ask whether it is a by-product of something that does play an appropriate or necessary part of our responsibility practices. If moral revulsion is a necessary or inescapable by-product of an imperfectly efficient system, then it would be safe from complaint for just this reason. Like the various forms of gas produced by human digestion, it might be the sort of thing that we put up with, manage, or ignore but whose elimination would (presumably) require drastic measures we are unwilling to undertake. My point is *not* that moral responsibility is like indigestion. Instead, my point is that one can allow that the justification of praise and blame might derive from the efficacy of those norms in influencing us, without thereby committing ourselves to the view that every instance, or even every type of characteristic emotional reaction, thereby contributes to influencing us in the appropriate way. Sometimes, counter-productivity is a necessary consequence of the most effective available system.

Consider a case in which a particular agent would be unmoved by praise, blame, or some display of responsibility attribution. On the traditional MI account, that agent could not be responsible for his or her actions. On the account I propose, whether the agent is morally responsible for his or her actions is not a function of that particular agent's susceptibility to influence in that particular circumstance, but rather a function of what the justified norms of moral influence say about the status of responsible agents in those contexts.[21] These norms (that is, the norms of responsibility) will be those norms—whatever they are—that are most effective at *collectively* influencing agents in the appropriate way. There is no reason to suppose that the contents of individual norms (as opposed to their justification and aim) or the practices that reflect those norms will themselves have a consequentialist character. On my account, the notion of moral influence is important as a higher-order phenomenon, one that describes the basis of justification for a network of practices, attitudes, and judgments. So, again, few if any of these first-order elements have a markedly consequentialist character.

4. Limits and Burdens: Norms, Modularity, Exclusivity, and Provisionality

My proposal has a somewhat different profile than traditional moral influence theories. So, before returning to the objections that felled the traditional moral influence account, I will briefly comment on some of the burdens and limits of the account I have advanced. In what follows, I also discuss a family of concerns this account may raise.

The Norms of Responsibility

I do not intend to say much about the content of the justified responsibility norms. Nevertheless, it should be clear that those norms, whatever they turn out to be, are structured by a range of concerns tied to the aim and justification of the responsibility system, the facts of human psychology, and the contexts in which the norms are applied. As we have seen, part of what the responsibility norms recommend in a given context will be upshots of larger theoretical demands, such as stability and psychological efficacy. Given the kinds of psychologies we have, and the time it takes to inculcate the relevant sensitivities and responsibility-characteristic reactions in people, a frequently fluctuating network of norms and practices would be a disaster for our ability to govern ourselves and others in compliance with justified responsibility norms.[22]

One might grant all of this yet still ask how we move from the justification of the responsibility system as a whole to the more particular justification for individual instances of praising and blaming. That a norm of coming to a complete stop at a traffic stop sign might be generally justified does not answer the question of whether adherence to it is justified in *this* case. It might, from a practical standpoint, be justified enough if I think I am in the ordinary case. But, if I am asking whether I am in that case, knowing that the norm is generally justified does not seem to settle the question. If the question is unsettled, it is not clear whether I should be particularly concerned with praise and blame.

In reply, I am inclined to think there is something right about the idea that, sometimes, we might not have reason to care about whether someone is praiseworthy or blameworthy, and that we might sometimes have sufficient reason to ignore the edicts of the responsibility norms. I will say more about this possibility in a bit. However, I think there is a second reply we should keep in mind, which is that, in the ordinary case, we have good pragmatic reasons for supposing that individual instances of praise and blame are justified in light of features about the general norm. It is not an infallible warrant, of course, but it is good enough for ordinary practices. That is, in the general case the facts about whether a given responsible agent deserves praise or blame for something will be settled, at least internal to the norms of moral responsibility, by what the norms say about cases *of that type*. The justification in the individual case follows this relation of fact to type. Since the norms are given in part by their general efficacy, we can expect that most cases will fall unproblematically under their scope, as instances where praise and blame plausibly play the right sorts of roles, given facts about the agents and their circumstances. These will be cases where praise and blame is justified or at least permitted as a tokening of those practices or judgments that are, in fact, justified in the ways I have described. This is where the possibility of indirect effects, necessary inefficacies, and our ordinary psychological messiness can do some work in helping us see how we ordinarily have a pragmatic warrant for thinking praise and blame is justified in the usual sorts of ways. Exceptions are possible, but there is no reason to think they will be the rule.[23]

Nevertheless, one might think we need more than a 'merely' pragmatic warrant for accepting that the norms of praise and blame are applicable. This demand strikes me as misguided, given that what is at stake is fundamentally practical in nature, especially given that we already have a story about whether or not an agent is really praiseworthy or blameworthy (i.e., we check to see what the norms say about cases of that type). What we ought to be looking for is a judgement good enough for guiding our actions and assessments of whether we should praise or blame. A pragmatic warrant gives us that. One might still object that the pragmatic warrant is attempting to track some prior and independent fact about whether praise or blame is really justified, and thus that independent normative status is still relevant. True enough. Settling these normative questions in that degree of metaphysical detail would be wonderful. It would give us more than a merely pragmatic warrant. However, obtaining such epistemic credentials is extraordinary difficult for almost *any* ordinary moral judgement. Yet, some judgement is still required of us on a regular basis in our often epistemically unextraordinary daily lives.

In sum, the most we can reasonably demand is a pragmatically warranted judgement about what seems to be the case, normatively speaking, in the circumstances we find ourselves. And, given that we accept that the norms of responsibility are justified as a whole because of how they contribute to our being moral considerations-sensitive agents, it seems to me that we will ordinarily have adequate grounds for believing that, in any typical case, these norms will likely apply.

One important class of cases where the usual warrant might be defeated are cases with agents in new, unusual, or particularly challenging contexts of action. In these cases, though, it looks like the right place to look for settling whether praise and blame is justified is not so much the theory of responsibility norms but, perhaps somewhat surprisingly, the account of responsible agency. Here is why: if the threat to a responsibility ascription is coming via some threat to the normal capacity to respond to moral considerations (as it seems to in cases of new, usual, or particularly challenging contexts of action), then the issue is how these concerns are accommodated internal to some account of the capacities required for being subject to the responsibility norms. In other words, we look to our theory of responsible agency. Once the capacity issue is settled, there may be some features of that agent that become relevant to the assignment of praise and blame (for example, perhaps the difficulty of responding to moral considerations is not high enough to render that agent a non-responsible one in that circumstance, but perhaps it is high enough to fund some degree of mitigation in blame). Nevertheless, the basic issue seems to be a challenge to the details of a theory of responsible agency more so than the details of a theory of the responsibility norms. So, at least internal to the norms of responsible agency it seems we can account for the justification of most concrete cases of praise and blame.

Nevertheless, there is something slippery about the question of justification in the particular case, something which may leave one with the sense that the account thus far fails to hit the mark. Perhaps the worry we should have is not with how we might settle the justification for particular cases of praise and blame, internal to the norms of responsibility. Instead, perhaps it is a kind of concern external to those norms. Since the norms receive their justification from more general facts about their efficacy in a community, one could worry that we can adopt a standpoint external to those norms. If so, then we might wonder whether there is justification for the enforcement of those norms in a particular case if we do not necessarily take ourselves to be bound by responsibility norms in general.

These issues are difficult, but it seems to me that there are two lines of reply.

First, I see no reason to suppose that a theory of responsibility must provide a decisive answer to the more basic normative question of what one ought to do, all things considered, even in cases where what is at stake is some matter of moral responsibility. This is because a theory of responsibility must be silent on those considerations whose origin or normative force places them external to the norms of responsibility. This more fundamental normative or deliberative task—settling the all-things-considered practical matter—even in cases concerning responsibility, is more clearly a task for a theory of normative ethics or a theory of practical reason. So, if we adopt a standpoint external to the norms of responsibility or from a standpoint that is skeptical of moral force in general, and ask how to close the gap between the justification for the responsibility norms and the justification of some particular instance of praise and blame, the question becomes uninteresting for a theory of responsibility. If the edicts of the responsibility norms are only inputs into some greater normative, practical calculation, a piece of deliberation where responsibility norms can be trumped by other concerns, then when we ask this question external to the norms of moral responsibility we cease to be talking about a question that must be answered by a theory of moral responsibility. At best, the responsibility norms identify the salient normative facts relevant to concerns of responsibility, but the ultimate question of whether one is justified in blaming, all things considered, is to be decided by appeal to considerations beyond the scope of this account.

A second line of response focuses on the externality issue in a different way. Given that the justification for any particular instance of blaming hinges on normative issues outside of merely a theory of responsibility, what may be at stake in asking the question (whether some particular case of blaming is justified) is only whether we can re-establish the warrant for adopting a standpoint internal to the norms of responsibility. Since this is warrant funded by some confidence in our moral assessments, the generally justified status of the norms, the ordinariness of our case and so on, all we can do is rehearse the reasons for caring about moral responsibility and working through the arguments for its importance. In re-establishing confidence in those things, we re-establish confidence in our judgements of concrete particular cases of responsibility ascriptions.

So, depending on how the question of justifying an individual ascription of responsibility is meant, there are four different replies that can be made on behalf of my account. First, we can appeal to the actual normative status of the agent under the justified norms of praise and blame, a normative status that is settled by, among other things, its falling (or failing to fall) under a type of action prescribed or proscribed by the norm governing that context, or alternately, by the action being conducive or antithetical to the aims of the responsibility system. Second, we can show how the motivating concern might really be about some other aspect of a theory of responsibility, namely, a theory of responsible agency. Third, if the question is one external to the norms of moral responsibility, we can reject the demand that a theory of responsibility must provide an 'all-in' account of the justification for a given instance of praise and blame. Fourth, we can endeavor to show that the usual pragmatic warrant for ascribing responsibility is in place.

Modularity

Although I have only gestured at a theory of responsible agency, gesturing at it may be sufficient to raise a different kind of question. In particular, we might wonder what constitutes those moral considerations to which I claim the responsibility system aims to make us more sensitive. The answer to this question is determined by the correct theory of normative ethics. So, the theory I propose is designed to illuminate something about the distinctive logic of moral responsibility in a way compatible with a wide range of (plausible) theories of normative ethics. This is what makes this account of responsibility *modular*. When integrated with different ethical theories, the account of moral considerations will change, but the basic structure of justification for the distinctive norms of moral responsibility will remain intact. Of course, if consequentialism is true, we should look to the true consequentialist theory of the good to inform our account of moral considerations. And, if Kantianism is true, moral considerations will be grounded in the categorical imperative. Since the moral influence theory is not intended to be an account of right action, but rather a broadly modular account of moral responsibility, you may fill in these details any way you like.[24]

Normative Exclusivity

A distinct but related feature of my view is that it is not *normatively exclusive*. A normatively exclusive account would maintain that this 'higher-order' moral influence account is the sole way of justifying our responsibility-characteristic practices and attitudes. However, I see no reason to dismiss the possibility that there may be other, perhaps imperfectly overlapping, alternative justifications that independently vindicate or modify some subset of our responsibility characteristic practices and attitudes. If any of our reactive attitudes or responsibility-characteristic practices have other sources of justification as well, then so much the better. For example, that the theory explains why moral praise of an agent is justified in terms of the effects of the relevant norm having currency in a society in no way precludes the possibility of other sources of justification for praising and blaming. Perhaps praise can be connected to the value of the will or the character trait that governs the action. Multiple overlapping justifications can coexist peacefully, and indeed, prove to be mutually supporting in our moral practices.

Provisionality

The rejection of normative exclusivity entails that my account is provisional. Discovery of additional, independent justifications for responsibility-characteristic practices and attitudes will potentially create conflicts where one justification counsels something that differs from the other.[25] For example, an alternative account of the justification of responsibility system fleshed out primarily in terms of a principle of fairness might, at various points, conflict with the account proposed here. We would then need to go in for further refinements in the account in light of this discovery. But these further developments would pose no serious difficulty for anyone already open to revisionism about moral responsibility—further revision is simply in keeping with the spirit of the project.

4. How this Account is Immune to Traditional Objections to Moral Influence Accounts

I have attempted to show how we might find some modest but not unimportant use for the idea of moral influence in the context of a theory of moral responsibility. Now I want to show how this limited deployment of the idea of moral influence might do its work without incurring the difficulties that beset traditional conceptions of moral influence.

Recall that the principal objections levied against traditional moral influence accounts were these: (1) the coarse-grainedness objection, that (a) MI theories cannot adequately distinguish between responsible and non-responsible agents and (b) that moral influence cannot be distinguished from other kinds of influence; (2) that MI theories cannot respect the distinction between being responsible and being appropriately held responsible; (3) that MI theories grossly mischaracterize how we praise and blame, and our concerns in doing so; (4) that MI theories cannot accommodate backward-looking moral concerns, and (5) that moral influence theories are inappropriately committed to a particular theory of normative ethics.

At this point it should be apparent that many of the objections simply do not apply to the more modest conception of moral influence I have been advancing in this paper. The details, however, are instructive.

Objections (1a) and (1b) are different aspects of the complaint that MI accounts are too course-grained in their handling of responsibility to be adequate as theories of responsibility. The objections were potent when directed at traditional accounts. However, they are clearly inapplicable to the less ambitious role to which MI has been constrained in my account. Regarding (1a), that an MI account cannot make suitable distinctions among agents, the answer is simple: My account does not rely upon MI to distinguish between responsible and non-responsible agents. That work is left to an account of responsible agency, which on my view is tied to a kind of moral considerations-mongering agency. Such an account, while compatible with the restricted use to which I put the idea of MI, does not itself depend on it. So, objection (1a) is defeated.

With respect to the other part of the coarse-grainedness complaint (1b), concerning MI's inability to distinguishing between kinds of influence, there is more to be said. Again, though, the reply hinges on the different labours assigned to the individual parts of a theory of responsibility. On the account I have offered, justified praise and blame involves the judgement that particular responsibility-characteristic attitudes (e.g., indignation) are licensed when directed at the target of evaluation.[26] In turn, this judgement presupposes that the evaluated agent is the right sort of agent to be a target for those reactions. So, on this account, the appropriateness of praise and blame is parasitic on the truth of the judgement that the target of praise and blame is a responsible agent. And, as we have seen, that is given by a theory of responsible agency and not a theory of the justification of the responsibility norms. In contrast, other forms of influencing the behaviour of agents have no such requirement on them, and indeed no such supposition ordinarily built into them. In influencing a household pet, there is (ordinarily) no judgement that the pet is a responsible agent. Hence, the form of regard expressed in distinctively moral praise and blame is not present. So, even if some of the *practices* of moral influence are superficially indistinguishable from non-moral influence, the underlying attitudes and judgments are distinct.

What this makes clear is that judgements of genuinely moral praiseworthiness and blameworthiness have a distinctive cognitive content to them, a content that makes error possible. We can believe that the relevant capacities are present when they are not and we can mistakenly suppose that they are absent when they are present. What makes a genuine ascription of responsibility *true* is (i) that the considered instance of moral influence corresponds to what a stable, justified responsibility system would prescribe or permit, given the facts about human psychology and given the aims of the responsibility system, and (ii) the agent is a responsible agent. So, even were we able to influence a cat's behaviour in light of expressions of responsibility-characteristic attitudes and practices (such as moral praising and blaming), it would nevertheless fail to be genuinely *moral* praise or blame unless the praiser or blamer also believed that cats were responsible agents. Presumably, we sometimes make errors in the case of humans (and maybe, sometimes in the case of cats). All this shows is that there are cases where an ascription of responsibility is mistaken, even if the praise and blame were real. And this is exactly what we should think. So, we have dispatched the second half of the coarse-grainedness objection.

The second major objection to traditional MI theories is that they cannot respect the important difference between whether someone is responsible and whether it is appropriate to hold someone responsible. Since agents can be responsible without it being appropriate to hold them responsible (recall the example of arbitrarily punishing only one person for his or her impolitic remark), any theory that collapses these distinct assessments fails to reflect an important feature of our thinking about responsibility. Traditional MI accounts appear to fail in just this way. They begin with an account of when we should hold someone responsible (when it is efficacious) and conclude that someone is responsible only when we should hold him or her responsible.

The account I have given permits a different response to this objection: whether someone *is* morally responsible depends on two things: (i) whether the evaluated agent is a responsible agent and (ii) what the justified norms of responsibility say about agents in cases of the considered type. As I suggested earlier, though, this is consistent with a view that emphasizes that we can ask questions external to the responsibility system. We can ask whether a responsible agent *ought* to be held morally responsible in light of, say, considerations of justice, benevolence, prudence, and so on—even if that agent is both a responsible agent and *in fact* morally responsible.[27]

One aspect of this picture is that it reflects a degree of modesty about the role that moral responsibility plays in our lives. It is important, but it does not and perhaps ought not override every other consideration in our lives. There are standpoints and concerns from which focusing on whether to praise or blame seems misplaced, even when there is a clear answer from the standpoint of moral responsibility. So, even if you thought that all-in moral considerations ought to be decisive in deliberation, it seems doubtful that the norms of moral responsibility (specifically, and by themselves) are the sort of thing that trump all other considerations.

Here the relevance of my account's modularity becomes salient. When engaged in the practice of praising and blaming, what we have reason to do depends in part on the resources of the background moral theory in which an account of moral responsibility is embedded. Although we can describe the general shape of a system of moral responsibility—its logic, as it were—particular cases will be decided by the integrated mesh of the norms of both responsibility and normative ethics.

To illustrate, consider the traditional consequentialist problem with scapegoating. Suppose that we learned that the most effective, stable set of responsibility practices involved blaming some group of people who had done no wrong. If the account I have suggested permits this, one might think this is a significant strike against it.

There are, I think, two different lines of response appropriate here, one turning on the particular details of the package of views I favour, the other deriving from more general features of a modular account. I will pursue these lines of response in turn.

First, although scapegoating could be a worry for some accounts that rely on the notion of MI for the justification of the responsibility norms, in this case it is precluded because of the account of responsible agency and the aims of the responsibility system I have invoked. The aim of the responsibility system, I have claimed, is to foster moral considerations-sensitive agency. I have also said that genuine judgements of praiseworthiness and blameworthiness contain a kind content to them, one where the agent is regarded as a moral considerations-responsive agent. One result of this picture is that ascriptions of moral responsibility require taking a kind of stance towards other agents, one with distinctive regard for the form of agency involved. Indeed, concern for that form of agency which has the capacity to be governed by moral considerations is really the point of the responsibility system. An important upshot of this is that the attitude of agential regard, characteristic of and partly constitutive of holding agents genuinely morally responsible, precludes scapegoating. Although I can only gesture at the argument, to scapegoat an individual or group would be to fail to regard those agents in a way that is concerned with respecting and fostering the form of agency with which the responsibility system is concerned. On the account I have endorsed of the aim of the responsibility system and the picture of responsible of agency, such an arrangement looks incoherent, requiring something that is fundamentally at odds with the conception of agential regard that constitutes the end of the practice, an end that structures the norms themselves. Thus, on the package of views I have been defending, scapegoating appears to be precluded.

While the upshot I have sketched is an upshot of my assumptions about the aim of the responsibility system and the kind of agency required for responsibility, there is a different kind of response available. This second response turns on the modularity of my account. That is, whether scapegoating is permitted partly depends on the features of the normative ethical theory with which the account of responsibility is integrated. For ethical theories that do not centrally countenance fairness and distributive justice, for instance, there will be few resources to rule out the permissibility of scapegoating. However, for theories that take these considerations to have substantial force, scapegoating will always already be precluded. For example, on many Kantian-inspired theories considerations of fairness, distributive justice, and respect for persons as ends would presumably always trump incentives to favour scapegoating. So, when a theory of responsibility is integrated with this sort of theory, scapegoating ceases to be a worry.[28]

As we have seen, there is a complex relationship between the norms of responsibility and a range of more general judgements, including all-things-considered judgements and (if they can come apart from all-things-considered judgements) judgements about what we have most moral reason to do. The chief lesson here has been that my modest use of the idea of moral influence does not trample the important distinction between when someone is responsible and when it is appropriate to hold him or her responsible. Indeed, this distinction can be rendered consistent with recognition that while a system of moral responsibility has something of an internal logic to its norms, it is nonetheless part of a broader system of normative ethics.

The last three objections can be dispatched fairly quickly. Consider objection (3) above: this is the objection that MI theories grossly misconstrue our responsibility practices, confusing a part of our responsibility practices (aiming to influence) with the entirety of our practices, and thus (perhaps) committing us to a perpetually therapeutic or 'detached' attitude towards praising and blaming. Whatever its virtue as a complaint against traditional MI accounts, it is clear that this objection finds no purchase against the account I have offered. In particular, nothing I have said presumes that all individual acts of praising and blaming are undertaken with an eye towards influence, or that those acts of praise and blame have a structure any different than the critic contends. What I have maintained is that the norms of responsibility are justified in light of the efficacy of those norms and the organic, diversely-motivated collection of practices that those norms give rise to. Indeed, it is plausible to think that for creatures with psychologies like ours, that efficacy precisely depends on our being interpersonally engaged, feeling gratitude, resentment, and the like. As I have already noted, a particular instance of praise or blame may not, in isolation, contribute to the aim of the responsibility system. Indeed, it may be a counterproductive but unavoidable aspect of a stable system (that is, it may be a necessary inefficacy).[29] So, adoption of a permanent therapeutic standpoint is neither obviously desirable nor necessary. We need not abandon a commitment to the reactive attitudes. Instead, they provide some of the most basic mechanisms by which justified norms of responsibility come to be effective in the world.

As this account is prescriptive and open to revision of our folk concepts, it is no objection to what I have said to argue that our current practices fail to be those that are maximally effective at fostering moral considerations-sensitive agency. First of all, it is not clear that what is required is the maximally effective set of possible practices, as opposed to a system that is sufficiently effective given the current costs of being more or equally effective. Second, the objection is surely right in its substance: it would be altogether stunning to learn that our exact norms and practices (messy as they are) happen to be exactly those that are best at fostering moral considerations-sensitive agency. I am inclined to think that in responsibility, as well as in many other domains, there is room for a kind of moral progress.

We can now consider the fourth objection, which holds that MI accounts have no place for responsibility-characteristic reactions (such as gratitude) that are backward-looking in their assessment. As P.F. Strawson pointed out, gratitude is among those attitudes that are particularly sensitive to the quality of will directed at us. That is, when others regard us with a good will, and in particular, when they act with good will towards us and we recognize it, we typically respond with gratitude. Gratitude thus helps mark recognition of a good will. Assuming a good will is at least sometimes reflective of moral considerations, it is reasonable to think that learning to track a good will can play a role in learning to track moral considerations. Perhaps more importantly, our reactions of gratitude can signal that we recognize that other agents are responding to what we regard as appropriately agency-guiding considerations. Of course, sometimes these considerations are extra- or non-moral, but inasmuch as gratitude reliably reflects appreciation of moral considerations-governed agency too, gratitude has all the license we could hope for it. Similar remarks hold for other backward-looking attitudes: as long as they plausibly play a role in the social and intrapersonal economy of governance by moral considerations, there is no objection here.[30]

Still, this talk of licensing may sound artificial. If gratitude is a response that is deeply and irrevocably part of our nature as social beings it may need no licensing. If so, then it is one of those elements of our psychological landscape around which any plausible theory of responsibility must be contoured. And in being such a thing, gratitude would be no difficulty for this account of responsibility.

We can now quickly dispatch the fifth and last objection, which holds that MI accounts are problematically committed on the matter of the correct theory of normative ethics. As we have already seeen, the present account is modular and does not rely on the truth of consequentialism. So, this criticism does not apply to my use of the idea of moral influence.

What all of this should show is that the difficulties that beset traditional accounts of moral influence have less to do with the idea of moral influence *per se* than they have to do with overplaying the proper scope of the idea of moral influence. If the role of moral influence is limited to the justificatory structure of the responsibility norms, then it can function as a sleek but powerful element in a larger theory of moral responsibility.

5. Desert and Depth

> The object of these commonplaces is to try to keep before our minds something it is easy to forget when we are engaged in philosophy, especially in our cool, contemporary style…
> —P.F. Strawson, “Freedom and Resentment”[31]

By way of conclusion, I wish to address a lingering concern that can arise whenever one discusses theories of moral influence.

In discussions about free will and moral responsibility, philosophers will sometimes say that if some or other argument for skepticism about moral responsibility is sound, then all that can be justified is some merely consequentialist conception of responsibility.[32] This account of responsibility is usually only gestured at, but it is invariably described as superficial, a kind of ersatz responsibility. This contrasts with notions of moral responsibility that are ‘deep’ or ‘ultimate’ in some desert-entailing way. The consequentialist conception philosophers usually seem to have in mind is the traditional moral influence account. And, presumably, the *de rigeur* tone of dismissiveness in these conversations reflects the failure of traditional moral influence accounts to satisfactorily address those objections I presented at the start of this paper.

I share the commonplace conviction that traditional moral influence accounts are inadequate theories of moral responsibility. Whatever the limitations of those accounts, my more modest use of the idea of moral influence is intended to be a part of an account of responsibility that justifies genuine desert-entailing attributions of responsibility. And, inasmuch as I understand what is meant by 'deep responsibility', this account is supposed to be part of a theory of what the bona fide, genuine, real, 'deep' sense of moral responsibility is, or at any rate what we *ought* to have in mind by that sense. The modest usage I make of the idea of moral influence does not preclude depth or desert-entailment in our ascriptions of responsibility, unless those things are meant in some question-begging way. Admittedly, it is not always clear to me what 'deep responsibility' comes to, supposing it is something more than a merely stipulative notion of some extraordinarily demanding conception of agency. Still, for all I have said, this account of the justification of the responsibility norms might be integrated with an account of responsible agency that requires whatever metaphysically robust conception of agency you like, up to and even beyond agent causation.

However, I do think that the account also undercuts some of the impetus for accounts of responsible agency that are more metaphysically extravagant (read: libertarian). Once it is evident that we can justify norms of responsibility along the lines I have described, and given that practices roughly like ours can make that justification viable on a range of accounts of responsible agency, the pressure for a libertarian conception of responsible agency begins to diminish. There is nothing in this account that suggests that we *need* agent causation, or indeterminism for that matter, to justify these norms and attendant practices of responsibility. Given that we do not need these things for the integrity of the bulk of our responsibility-characteristic practices, attitudes and judgments, it is not clear what exactly turns on requiring these further conditions.[33]

Here, though, is precisely where talk of depth or ultimacy sometimes re-emerges. Perhaps there is something special, to be desired, or valued in moral responsibility that cannot be gotten on the account I have given. What that is, however, needs to be brought to light. Invocations of depth too often obscure more than they reveal, masking what *ought* to be our fundamental interest here. And, I take it, what we ought to be concerned with is the answer to this question: 'What are the conditions under which we are entitled to treat others better and worse, where that involves merited praise and blame, reward and sanction, and so on?' Answering this question requires a theory. However, there is no reason to suppose that this theory will perfectly enshrine our pre-philosophical intuitions about moral responsibility. Part of the point of theorizing just is to break new ground, to potentially learn that the world is somewhat different than we anticipated. This holds true in the case of deep, desert-entailing responsibility as much as it does in the case of human rights, constitutional government, and astronomy. If I am right, the conception of moral responsibility I have been defending may be exactly what we are looking for, even if it wasn't exactly what we had in mind.[34]

Notes

[1] P.F. Strawson, "Freedom and Resentment," *Proceedings of the British Academy* 48 (1962):1-25, reprinted in Gary Watson, *Free Will*, second edition (New York: Oxford University Press, 2003), pp.72-93. Quotation from page 74, emphasis in original.

[2] T.M. Scanlon, "The Significance of Choice," in Sterling M. McMurrin (ed.), *The Tanner Lectures on Human Values*, vol. 8 (Salt Lake City, UT: University of Utah, 1988), p.159.

[3] The classic statement of a theory of this sort is Moritz Schlick's in ch. 7 of Moritz Schlick, *The Problems of Ethics*, trans. D. Rynin (New York: Prentice Hall, 1939), reprinted as "When is Man Responsible?" in Bernard Berofsky, *Free Will and Determinism* (New York: Harper & Row, 1966). An interesting, and somewhat revisionist twist to the view is given by J.J.C. Smart, "Free Will, Praise, and Blame," *Mind* 70 (1961): 291-306. Richard Arneson has recently offered a rehabilitation of Smart's account that is congenial to some of the points I make here. See Richard J. Arneson, "The Smart Theory of Moral Responsibility and Desert," in Serena Olsaretti (ed.), *Desert and Justice* (New York: Oxford University Press, 2003), pp.233-58.

[4] In the free will literature, important statements include: P.F. Strawson, "Freedom and Resentment"; Jonathan Bennett, "Accountability," in Zak Van Straaten (ed.), *Philosophical Subjects* (New York: Clarendon Press, 1980), pp.14-47; T.M. Scanlon, "The Significance of Choice"; and R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press, 1994), pp.54-59. To find the most recent notable defence of it, you have to go back almost twenty-five years to Daniel Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: MIT Press, 1984). Significantly, reviewers with widely divergent estimations of the book generally agreed that Dennett's defense of moral influence was unsatisfactory. See Gary Watson, "Review of *Elbow Room*," *The Journal of Philosophy* 83 (1986): 517-22, and Gerald Dworkin, "Review of *Elbow Room*," *Ethics* 96 (1986): 423-25.

[5] For objections in the spirit of what I have been discussing, see C.A. Campbell, "Is 'Free Will' a Pseudo-Problem?," *Mind* 60 (1951): 447, and P.F. Strawson's "Freedom and Resentment". It is, of course, open to the moral influence theorist to insist that we are better off without these distinctions, but it is also clear that in

going this route we would be abandoning a substantial part of our given conceptual furniture associated with moral responsibility.

[6] If I understand him properly, Scanlon seems to have something like this in mind when he claims that "the theory appears to conflate the question of whether moral judgment is applicable and the question of whether it should be *expressed* (in particular, expressed to the agent)." See Scanlon, "The Significance of Choice," p.159 (emphasis in original).

[7] Bennett, "Accountability," p.20. The same criticism is made by Strawson in "Freedom and Resentment". Similarly, we should expect puzzlement if it turned out that one could always express gratitude while being in a perfectly stormy frame of mind about the considered person.

[8] See Dworkin, "Review of *Elbow Room*," p.424: "Any attempt to forge as close a link between responsibility and modifiability…ignores those ascriptions of responsibility which are not oriented toward the future but are, so to speak, for the record. And since they are for the record, justice requires that we pay attention only to the details of a person's circumstances, and not to what is true in general or true of individuals very similar to her." Similar objections can be found in Wallace, *Responsibility and the Moral Sentiments* pp.56-57, Robert Kane, *The Significance of Free Will* (New York: Oxford University Press, 1996), p.83, and Campbell, "Is 'Free Will' a Pseudo-Problem?," p.447.

[9] Henry Sidgwick, *The Methods of Ethics*, 7th edition (Indianapolis, IN: Hackett, 1981), p.284.

[10] See John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, *Four Views on Free Will* (Malden, MA: Blackwell, 2007), ch. 4.

[11] John Martin Fischer and Alan Hájek both encouraged me to address this concern.

[12] Although the individual sub-theories of a theory of moral responsibility might be conceived of and treated independently, I am inclined to think the most promising way to develop an account will be one whose parts are interrelated and ordered by some general conception of what responsibility is ultimately about. In what follows, I hope to show that even if this is true, there is good reason to keep clear the distinct explanatory burdens of the various parts of a Grand Theory of Responsibility.

[13] Conceptions of the responsibility system that emphasize the enforcement model or aim at achieving some more particular 'overall best' result will likely run afoul of Kantian concerns about using people merely as means to whatever end is specified by the alternative conception. This concern may apply to Arneson's recasting of Smart's theory of responsibility in Arneson, "The Smart Theory of Moral Responsibility and Desert".

[14] One might wonder whether there is need for a separate theory of responsible agency, or if there is need for one, why it should not just fall out of a theory of moral responsibility. There are two things to be said here. First, although a theory of responsible agency and a theory of the content of the norms of responsibility are presumably importantly interrelated, it does seem possible that the considerations that govern who is subject to the norms of responsibility should be of a very

different kind than those that govern the content of the responsibility. Or, to borrow some language from the Strawsonian tradition (and, in particular, Gary Watson's discussion in "Responsibility and the Limits of Evil," in his *Agency and Answerability: Selected Essays*, New York: Oxford University Press, 2004, pp.219-59), a theory of exemptions (which concern agents as a whole) might operate on very different principles than a theory of excuses (which concern actions and the agent's relation to them). So, theorizing should respect this possibility by carving up these domains accordingly. Second, albeit relatedly, there are independent reasons for thinking that responsible agency has a particular value discrete from its role in moral responsibility. For example, one might accept a Kantian story about the intrinsic value and dignity of the form of agency identified as responsible agency. Or, one might think this distinction is useful as a way to characterize the moral agent/moral patient distinction, which concerns the kinds of entities and interests we need to respect and how we weigh them.

[15] See "Building a Better Beast" (in progress) and "Situationism and Responsibility" (in progress). My account derives much of its inspiration from reasons-oriented views that have been developed by numerous figures prominent in the literature on free will and moral responsibility, including John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge University Press, 1998), Wallace, *Responsibility and the Moral Sentiments*, and Susan Wolf, *Freedom within Reason* (New York: Oxford University Press, 1990). See also accounts by Michael McKenna, "The Limits of Evil and the Role of Moral Address," *Journal of Ethics* 2 (1998): 123-42, and Nomy Arpaly, *Unprincipled Virtue* (New York: Oxford University Press, 2003). There are important differences between these accounts and my account, of course. However, the particular details are immaterial for present purposes.

[16] For ease of exposition, I am putting to the side some complexities concerning cases where an agent voluntarily and intentionally undermines his or her considerations-sensitive capacities, or where an agent has been manipulated into having this capacity. For my account of how these cases are to be handled, see "On the Importance of History for Responsible Agency," *Philosophical Studies* 127 (2006): 351-82.

[17] Elements of this basic picture have a long history to them, stretching back at least to accounts of the internalization of norms found in Essay 2 of Friedrich Nietzsche's *Genealogy of Morality* (Indianapolis, IN: Hackett, 1998) and Sigmund Freud, *Civilization and Its Discontents* (New York: Norton, 1989). I do not share all the details with either of these accounts, but they offer suggestive accounts of how moral norms come to shape the psychology of agents even under conditions where no external threat is present.

[18] Compare the indirect forms of consequentialism presented in Robert Merrihew Adams, "Motive Utilitarianism," *Journal of Philosophy* 73 (1976): 467-81, and Arneson, "The Smart Theory of Moral Responsibility and Desert".

[19] Two points: First, you could think that there are two normative standards that are relevant here, where the first describes the 'normatively best' or ideal theory,

and the second describes what norms are possible for creatures like us to satisfy. Resentment could thus turn out to be unjustified in the ideal sense, but justified in the 'best we can do' non-ideal sense. Obviously, it is the second standard that I am concerned to meet. The second general point to be made here is that the potential discovery that the operations of gratitude and the other reactive attitudes are generally inescapable consequences of our psychology is compatible with those inescapable features having indirect benefits.

[20] This account of the justification of the responsibility norms has some parallels with the account of the justification of norms given in David Copp, *Morality, Normativity, and Society* (New York: Oxford University Press, 1995).

[21] This is thus an account that provides an explanation of what it is to *be* responsible, and perhaps by extension, an account that permits us to speak about whether or not someone is, in fact, responsible. Throughout, I will speak of normative facts. This is not intended to reflect a principled stand on familiar debates concerning moral realism, but is instead used purely for facility of expression. If some version of noncognitivism is true, then there should be some way to smoothly translate talk of purportedly normative facts into non-normative vocabulary.

[22] These considerations also explain why an analogue of a traditional objection to rule-utilitarianism does not get much traction here: one familiar objection to rule-utilitarianism is the charge that it collapses into act-utilitarianism because the best system of rules would be the one that has rules about individual cases. In at least the case of moral responsibility, the second-order moral influence theory is buttressed against such a collapse by the limitations of our psychologies, including the length of time it takes to develop and refine moral attitudes, the flexibility of our attitudes, the cognitive burden involved in assessing responsibility, and the overarching need to have a stable and efficacious system of influence. Collectively, these considerations will tend to weigh against something like act- or token-specific norms of responsibility.

[23] One might wonder what happens if the marginal cases are frequent enough. But there is something incoherent about the worry—what justifies the whole of the system is precisely that it gets the right results in the majority of cases. It might be conceivable that there is a world in which there is no system of responsibility-characteristic practices and attitudes that jointly generate justification for praising and blaming. Such a case would provide grounds for skepticism about the whole project of moral responsibility, funding a kind of skepticism about moral responsibility that would respect its conceptual and practical role in a way that most prominent forms of responsibility skepticism do not. But the circumstances of such a case seem sufficiently remote from our world so as to be of no concern.

[24] I am supposing that many moral notions could survive in the absence of libertarian freedom. Certainly many, if not most, philosophers working in normative ethics (including Kantian ethics) seem to accept something like this point, and many theories of justice seem to operate without presuming a notion of libertarian agency (though see Samuel Scheffler, "Responsibility, Reactive Attitudes, and Liberalism in Philosophy and Politics," *Philosophy & Public Affairs*

21 (1992): 299-324 for some complexities). So, even if you believe that some moral notions are jeopardized by the absence of libertarian freedom, surely not all moral notions are. And this paper is something of an argument for how core moral notions relevant to moral responsibility can be justified independent of libertarian pictures of agency.

[25] Other reasons for acknowledging the provisional nature of the account include the possibility that, as our practices and the attendant psychology adjust to different circumstances, what is justified will change.

[26] Here I focus on judgements of blameworthiness or praiseworthiness. The consequent blaming or praising attitude might be stymied for any number of reasons, so one might think someone is blameworthy for a particular action without in fact actively blaming that person.

[27] An example may help illustrate the point. Whether someone is a citizen is a matter of the laws of a particular country. This is an assessment that is made internal to a nation's system of legal norms. This legal fact does not mean that considerations external to the legal code cannot influence our judgement of whether it makes sense to treat someone as a citizen. We might well think Rogelio isn't, in fact, a citizen but still think that he or she ought to be treated as a citizen. Non-legal cases are possible, too. Peter may well be a jerk, but for a variety of reasons (perhaps we need his cooperation in some endeavour) we may decide that we will not treat him as a jerk. Similarly, whether someone is morally responsible depends, in part, on what the norms of the responsibility systems say about the agent or the action. Whether it is appropriate, fair, expedient, sensible, etc. to treat someone as responsible is a further, distinct issue.

[28] There are at least two different ways to conceive of the relationship between responsibility norms and the norms of normative ethics. On the first, you could hold that the content of the responsibility norms I have been describing is incomplete until it is filled in or provided with additional content by the norms of normative ethics. On the second way of conceiving the issue, the content of the norms of responsibility is complete and independent of normative ethics, but would be constrained by the norms of normative ethics in the way that, say, some egoistic considerations (whose content is often free of the flavour of morality) are constrained or trumped by moral considerations. I conceive of things in the former way, as I suspect that the latter way of picturing things is an artifact of artificially separating responsibility from the rest of normative ethics.

[29] A reminder by way of example of the basic argument for this view: Suppose the aim of philosophy is something like the identification of truths. We might plausibly suppose the ethos required to effectively pursue truth in philosophy may support and encourage the study of subjects that will not and perhaps could not discover truths. The norms required to make a practice effective at its aim may permit and encourage things that do not themselves contribute to that effectiveness. Moreover, it may turn out that some of these non-truth-conductive activities may have some feature that is independently valuable, perhaps more so than the aim of the activity that permitted it in the first place. Even if I thought that philosophy was about the pursuit of truth and that this permits any idea, regardless how crazy it

may be, to be published as long as it has a suitable defence, I could still think that a lot of great (and false) philosophy is valuable because it is consistent with values of originality, autonomy, self-expression, creativity, being inspirational, and so on. And, I might think any of these things are as valuable as the truths that philosophers are wont to discover, even though these other goals sometimes get in the way of finding the truth. As in the case of philosophy, realization that there are multiple valuable ends to which something may serve allows a degree of permissiveness in our practices.

[30] As a matter of theory, I suppose it is possible that the experience of gratitude would interfere with the operations of an effective and stable system of practices. Still, it seems extremely unlikely. However, were this unlikely scenario true, all it might show is that gratitude is an unnecessary evil in the technical sense I have used—the kind of thing that we might wish we could get along without, but which we cannot. But suppose we learned that gratitude was indeed one of the things that we could give up if we had sufficient reason to do so. Perhaps we could train ourselves and our children to never experience gratitude, and perhaps we could restructure a large sector of our interpersonal practices to reflect the expulsion of gratitude from our psychological economies. Would this change anything? I doubt it. Even if we discovered that we could give up gratitude, it is not clear that we would have any reason to. Even if it did not contribute to the aims of the responsibility system, as long as it did not interfere with them, there would be no incentive to get rid of it. But, if it did interfere with the aims of the responsibility system, as seems very unlikely, it would have to turn out that there is *no* independent justification for gratitude. Among other things, we would have to know that gratitude is not necessary for some other aspect of our lives that is valuable to us, and that no considerations external to responsibility favour it. The same goes for any other backward-looking attitudes.

[31] Strawson, “Freedom and Resentment,” in Watson, *Free Will*, p.77, emphasis in original.

[32] For my own part, I think extant arguments for (purportedly) skepticism about moral responsibility typically fail to be that and are instead, at best, arguments for skepticism about the proponent’s conception of what moral responsibility (or free will) comes to. I make this argument in more detail in Manuel Vargas “Libertarianism and Skepticism About Free Will: Some Arguments Against Both,” *Philosophical Topics* 32 (2004): 403-26, and in
Fischer *et al.*, *Four Views*, pp.145-48, 210-211.

[33] One might reply ‘free will’. However, it becomes notoriously difficult to say what free will is or why we should care about it, if it is not the control condition on moral responsibility.

[34] This paper has been in development for some time, and undergone a number of significant changes over the years in large part because of the numerous colleagues and friends who patiently explained to me what was wrong with it. Among those I should particularly acknowledge for their help in improving my doubtlessly still-flawed ideas connected to this paper are Michael Bratman, Andrei Buckareff, Meir Dan-Cohen, Keith Dromm, John Martin Fischer, Peter Graham, Alan ‘H-Bomb’

Hájek, Caspar Hare, Elizabeth Harman, Chris Hitchcock, Agnieszka Jaworska, Shaun Nichols, Dana Nelkin, Derk Pereboom, Katie Richardson, Yonatan Shemmer, Dan Speak, Ken Stalzer, Gary Watson, and audiences at the 2005 Pacific APA, UC Riverside, the SENT Workshop at Stanford, the LSU Conference on Applied and Theoretical Ethics, and the *Utilitarianism 2003* Conference in Lisbon, Portugal. Thanks also to members of seminars at the University of San Francisco and the California Institute of Technology for discussion of this paper. While working on this paper, I also received generous financial support from the Mortimer Fleishhacker Family Endowment for Philosophy at USF and the California Institute of Technology.