The Philosophical Quarterly Vol. 69, No. 276 *2019*
ISSN 0031-8094 doi: 10.1093/pq/pqy062

INSTRUMENTALISM ABOUT MORAL RESPONSIBILITY REVISITED

By Anneli Jefferson

I defend an instrumentalist account of moral responsibility and adopt Manuel Vargas' idea that our responsibility practices are justified by their effects. However, whereas Vargas gives an independent account of morally responsible agency, on my account, responsible agency is defined as the susceptibility to developing and maintaining moral agency through being held responsible. I show that the instrumentalism I propose can avoid some problems more crude forms of instrumentalism encounter by adopting aspects of Strawsonian accounts. I then show the implications for our understanding of responsibility: my account requires us to adopt a graded notion of responsibility and accept the claim that certain individuals may not be responsible because they are not susceptible to being influenced by our moral responsibility practices. Finally, I discuss whether the account is committed to allowing the instrumentalization of non-responsible individuals in cases where blaming them may benefit others' moral agency.

Keywords: moral responsibility, consequentialism, reactive attitudes, moral influence, instrumentalization.

When we think about responsibility, we always have one eye to the future and one eye to the past. We look at agents' past behaviour and intentions to establish desert, and we look to the consequences it is appropriate to visit on them in the future. In his seminal 1961 paper 'Free will, praise and blame', Smart put forward an instrumentalist account of what it means to be morally responsible. In a nutshell, Smart's claim is that the ascription of moral responsibility to an agent and praise or blame are justified if they have the effect of improving the agent's behaviour.

> If Tommy is sufficiently stupid, then it does not matter whether he is exposed to temptation or not exposed to temptation, threatened or not threatened, cajoled or not cajoled. When his negligence is found out, he is not made less likely to repeat it by threats, promises, or punishments. On the other hand, the lazy boy can be influenced in such ways. Whether he does his homework or not is perhaps solely the outcome of environment, but one part of the environment is the threatening schoolmaster. Threats and promises, punishments and rewards, the ascription and the non-ascription of

> responsibility, have therefore a clear pragmatic justification which is quite consistent with wholehearted belief in metaphysical determinism. (Smart 1961: 302)

Using the contrast between the stupid boy and the lazy boy as illustration, Smart argued that all it takes to be a responsible agent is to be an agent who can be influenced by praise and blame. A consequentialist account of responsibility thus has the benefit of making our practices of holding responsible and their justification central to the concept of moral responsibility, while avoiding controversial and, in Smart's view, incoherent metaphysical commitments to indeterministic free will. Others, however, have been less impressed. Smart's theory has come under fire for misdescribing our practices of holding each other responsible and inadequately capturing our understanding of moral responsibility. Critics object that Smart's proposal neither captures what we take ourselves to be doing when we make responsibility judgements and hold each other responsible nor does it capture what we *should* be doing. The latter point is important as arguably, most, if not all, accounts of responsibility have a revisionist element, but it is incumbent on proponents of an account to show that the revisions they propose improve the account.[1]

Recently, there has been a revival of (somewhat) instrumentalist accounts of moral responsibility, or, as Manuel Vargas calls them, 'moral influence theories'. Vargas (2008, 2013) has endorsed a partially instrumentalist account of moral responsibility, as has McGeer (2015), while Arneson (2003) has shown how Smart's theory might be revised to address many of the criticisms people have raised against it.

In this paper, I defend an instrumentalist account of moral responsibility and argue that it can meet the objections commonly raised against traditional instrumentalist accounts, either by showing that the objection in question doesn't arise on the revised account or that what objectors take issue with is not actually a problem. I argue that, not only are our responsibility practices justified by their effects, but that susceptibility to developing and maintaining moral agency through being held responsible is what makes for responsible agency. In this, I depart from Vargas, who sees a more limited role for instrumentalism as justifying our practices of holding each other responsible and proposes a separate account of morally responsible agency. Taking up arguments from McGeer and Arneson, I briefly show that instrumentalism can adopt aspects of more Strawsonian accounts, even though they are normally considered to be very different types of account (Shoemaker 2015; Wallace 1994). I then show important implications of my view: we can hold on to the claim that sometimes an individual can be a responsible agent without blame being useful or beneficial in a specific situation. However, the idea that

[1] See Elzein (2013) and McCormick (2013) for a criticism of Vargas' revisionist account as failing to meet this criterion.

non-responsible individuals, e.g. children, can still be an appropriate target of our responsibility practices is incompatible with my account. I argue that we should adopt a graded notion of being responsible and holding responsible according to which children are partially responsible and this is reflected in our blaming and praising practices. Finally, I discuss a potential problem for my account, which is that the justification of our moral responsibility practices also justifies holding the non-responsible responsible, if doing so benefits *others*.

In Section I, I outline the role consequentialism should play in an account of moral responsibility. I then show how the account deals with some of the standard objections to instrumentalism about moral responsibility and how it can be brought closer to a Strawsonian account in Section II. In Section III, I address important implications of my account. One consequence of the link between susceptibility to influence and moral responsibility is that individuals whose moral agency cannot be developed through our practices of holding responsible are not morally responsible. This marks an important difference to Vargas' account.

A further consequence is that my account does not allow for cases where the judgement whether an individual is a responsible agent and the judgement whether they are a suitable target for our responsibility practices come apart. Many see children as an example for this kind of case. I argue that we should accept this result and show how to make sense of the case of children. Finally, I discuss a morally problematic gap that can open up between being a responsible agent and being an appropriate target for our responsibility practices on my account. Indirect instrumentalist accounts run the risk of instrumentalizing individuals because they do not justify holding responsible merely with a view to the agent who is being praised or blamed, but also through the effects on others' moral agency. This is a troubling problem, but I will argue that in practice, others' moral agency does not benefit from blaming the non-responsible.

I. WHAT ROLE FOR CONSEQUENTIALISM?

Instrumentalist accounts of moral responsibility see the practice of holding accountable as central both to our concept of responsibility and to the justification of our practices of holding each other responsible. While many authors concede that punishment and reward should be sensitive to the effects that can be achieved by punishing and rewarding, they do not take the putative benefits of holding individuals to account and sanctioning them to lie at the core of moral responsibility. Accordingly, Smart's instrumentalist account is frequently criticized for missing the point, being overly simplistic, exclusively behaviourally oriented and insensitive to the psychology of our blaming and praising behaviour.

Like other newer instrumentalist accounts, I want to retain the close tie between responsibility and accountability to others, as well as the idea that the primary justification for holding others responsible is that this will lead to moral improvement. I take consequentialist considerations to be central to the justification of our practices of holding each other responsible. They go beyond a moral assessment of the quality of the action to a justification of our decision to hold a person to account for what they did. However, Smart's account is indeed unsatisfactory because of its narrow focus on affecting changes in behaviour. The aim of our practices of holding responsible should be conceived more broadly. One idea which can be found in the work of Arneson (2003), Vargas (2013) and McGeer (2015) is that our practices of holding each other responsible serve the aim of fostering moral agency or moral reasons responsiveness. On these accounts, the purpose of people's practices of holding each other responsible is to help them develop their sensitivity to moral reasons for action and their willingness to behave morally.

We should adopt this modification. Rather than merely aiming at compliance with moral demands without concern for the inner states of the creature one is trying to get to comply, the aim should be to change both moral thought *and* action. Furthermore, developing moral agency provides a general, not a local, justification for our moral responsibility practices (Vargas 2013). This means that it is not required that every single instance of holding responsible needs to have a positive effect in order to be justified. Rather, if our practices overall have the effect of making us into more moral people than alternative practices would, then they are justified. While this justification for our responsibility practices ultimately appeals to their consequences, it takes a detour via human psychology and claims that the justification for our current practices of ascribing responsibility is that they are suited to make creatures like us better and more responsive to moral considerations. The role of instrumentalism is therefore more indirect than in Smart's classic account, it is the general tendency to promote moral agency that justifies our responsibility practices, not the effect on one specific agent in one specific situation.

However, a justification of our responsibility practices is not yet an account of what makes a responsible agent. By 'responsible agent', I mean the kind of agent who can, in principle, be praise- and blameworthy and who is the fitting target of our responsibility practices. Responsible agents are the right kind of creatures to be part of the responsibility game; blame- or praiseworthiness for specific actions is a further issue.

What has been presented so far could just be an account of blame and punishment, praise and reward, not of what people need to be like in order to be responsible agents. So, on a hybrid theory of moral responsibility, we could establish responsibility in some independent way but then apply a purely consequentialist rationale when we decide what to do with the people we have found to be responsible. In fact, Vargas separates the justification of

our moral responsibility practices from the account of what makes agents morally responsible. He provides an independent account of moral agency, which specifies the capacities an agent needs to possess to count as morally responsible, as well as a theory of the responsibility norms which specify when moral praise, blame, etc. are justified.

The approach I take gives a more central role to moral influence. Consequences are not only relevant to justifying our practices of holding each other responsible. Rather, the fact that their moral agency is susceptible to being fostered and scaffolded by being held responsible is also what makes an agent a morally responsible agent, because it is what makes them the right kind of target for these practices. This allows us to retain the close link between the purpose of holding responsible and being responsible so central to Smart's account. This general account gives us an answer to the question: What kind of creatures are morally responsible? It does not tell us which actions are blameworthy or praiseworthy and in what situations (generally) responsible agents are excused for specific actions. Answering these questions is not the goal of the paper. My account can be seen as belonging to the same family as McGeer's (2014, 2015) and McGeer & Petitt's (2015) position regarding the relationship between moral influence and moral responsibility. However, because I endorse the view that the effect of our moral responsibility practices on *all* participants contributes to justifying them, it seems that the justification of our practices and the verdict whether an individual is morally responsible can still come apart. The worry is that this happens if blaming the non-responsible fosters moral responsibility within the moral community as a whole. I address this problem and offer a solution to it in Section III.2. I also develop the implications of an instrumentalist account further by proposing a graded account of responsibility.[2]

I will now flesh my account out further by showing how it answers standard objections to Smart's original account and how it relates to a family of positions that is often seen as incompatible with instrumentalism, Strawsonian or reactive attitude accounts of moral responsibility.

II. PUTTING MEAT ON THE BONES

There are a number of related objections to traditional instrumentalist accounts, many of which target the fact that a purely instrumentalist account of moral responsibility does not capture the phenomenology of our practices of holding responsible. Proponents of reactive attitude accounts of moral

[2] McGeer (2015) does also give a graded account of moral responsibility in terms of different levels of reasons responsiveness, but not as a response to the worry about a disconnect between our practices of holding responsible and the responsibility of the agent.

responsibility in particular have pressed this point (Shoemaker 2015; Wallace 1994). But in some ways, reactive attitude accounts are quite similar to instrumentalist accounts, as they, too, take *holding* responsible as central in explaining what it is to be responsible and are characterized as forward looking.[3] Furthermore, like Smart and other instrumentalists, reactive attitude theorists claim to have found a notion of moral responsibility which is not hostage to the metaphysics of free will, but based on our actual moral responsibility practices. Some authors even interpret Strawson as a consequentialist, pointing to the fact that he makes much of the desirability of our moral responsibility practices, irrespective of their metaphysical warrant (Arpaly 2006). McGeer provides a (controversial) interpretation of Strawson as an indirect consequentialist in her paper 'P.F. Strawson's Consequentialism' (2014). Nevertheless, she rejects Smart's account as exemplifying a notion of blame as brute deterrent intervention. But this criticism can be avoided if Smart's consequentialism is modified to aim at morally responsible agency. If we take this step, we can accommodate a lot of the elements important to reactive attitude theorists.

One reason why reactive attitude theorists are dismissive of instrumentalist accounts is because they believe that purely instrumentalist views do not do justice to the interpersonal involvement that is so central to our actual reactions of praise and blame. Reactive attitudes involve strong emotionally charged attitudes such as resentment, gratitude, but also guilt or regret from the first person perspective. Instrumentalist accounts, so the criticism goes, get the phenomenology completely wrong. Strawson makes the following complaint regarding consequentialist accounts of responsibility: 'But the only reason you have given for the practices of moral condemnation and punishment (. . .) is the efficacy of these practices in regulating behaviour in socially desirable ways. But this is not a sufficient basis, it is not even the right sort of basis, for these practices as we understand them' (2008: 4). Behaviour management, so the criticism goes, is not the same as holding responsible. When we punish or reward a dog in order to discourage or encourage certain behaviour, we are not holding it responsible. On Smart's account, it becomes hard to draw the distinction between holding responsible and operant conditioning.

By targeting responsible agency and responsiveness to moral considerations rather than morally desirable behaviour *simpliciter*, many of the implausible features of instrumentalist accounts are avoided. The broader focus on moral agency, instead of behaviour alone, allows the new moral influence theorists to do justice to the psychology of our moral responsibility behaviour in a way that Smart was unwilling to do.

[3] Though this is often interpreted as response dependence about moral responsibility, whereby our typical reactive attitudes somehow underwrite an individual's praise or blameworthiness. This rather mysterious thesis has recently been worked out and defended by Shoemaker (2017).

Smart (1961) wanted to give up any and all beliefs and emotions that he took to be tainted by residual belief in libertarian free will. Arneson (2003) holds that reactive attitudes may be instrumentally justified even if it turns out that Smart is right in claiming that they involve some mistaken judgements regarding free will. There is no reason why an instrumentalist account should not be able to take on board reactive attitudes, if these are central to reinforcing individuals' moral commitments and behaviour. Furthermore, considering that the stated aim of a reactive attitude account is to show how we can have responsibility without a commitment to libertarian free will, it is at least not obvious that emotionally charged blaming and praising behaviour is *ipso facto* committed to a libertarian notion of free will. Indeed, if the goal is to develop moral agency, then keeping reactive attitudes such as resentment or guilt in place may well be a good idea, as the emotional component involved in both our other-directed and self-directed acts of holding responsible will make praiseworthy acts seem more desirable and blameworthy ones more undesirable. In a recent paper, McGeer & Pettit (2015) also stress the importance of our consciousness that our behaviour will be morally judged by others for developing moral competence and reasons responsiveness. They argue that the real or imagined reactions of others to our behaviour and our need to justify ourselves to others help to scaffold our moral agency.

We can see how an instrumentalist account enriched by reactive attitudes might work by drawing an analogy to Railton's (1984) self-effacing consequentialism.[4] Even if the ultimate goal of our responsibility practices is to change behaviour and moral sensibilities for the better, an indirect approach may be better suited to achieving this aim. We don't have to have the goal of improving behaviour in mind every time we make a judgement about blameworthiness or blame somebody. If we don't keep our eye on a consequentialist goal in every single instance, this may have better effects on the development of moral agency and it will also allow us to have more personally involved relationships, a worthwhile outcome in its own right. Whether self-effacing instrumentalism leads to better results is in the end an empirical question,[5] but it is certainly more compatible with close interpersonal relationships than an explicitly instrumentalist, detached approach.

What I have said goes some way towards defending the consequentialist account from Strawson's criticism of 'one-eyed utilitarianism' (2008: 25). According to Strawson, the consequentialist (in his words, the optimist) 'seeks to find an adequate basis for certain social practices in calculated consequences, and loses sight (perhaps wishes to lose sight) of the human attitudes of which these practices are, in part, expression' (2008: 25). An account which explicitly makes room for the value of interpersonal relationships and does not characterize

[4] Thanks to Christopher Bennett for drawing this to my attention.
[5] There is also some discussion as to whether it is psychologically viable, see Doris (2015).

our blaming and praising as directly instrumental does not fall prey to Strawson's calculation objection. It can thus retain the phenomenology of our moral responsibility practices by showing that, when we judge someone to be blameworthy and blame them, we need not see this as an educational exercise.

However, some Strawsonians will still see this kind of instrumentalism as falling prey to a wrong-kinds-of-reasons objections, because they believe that what makes our blame and praise appropriate gets decided by looking backward at the agent's conduct and motivations, not by looking forward at the likely results of our reaction to it; even if we only consider effects for the purposes of the overall justification of our practices of holding responsible, and not in the heat of the moment. I agree that the *evaluation of the agent's action* gets settled by looking backwards. But once we have evaluated the agent's action and decided whether the agent was being wilfully immoral (acted with ill will or lack of good will), we still need to decide what the appropriate reaction to such behaviour is. The justification for what reaction is appropriate will lie in the effect on our interpersonal relationships and the development of moral agency. Importantly, this means that the justifiability of our reactive attitudes and our praising and blaming practices will depend on whether they do in fact foster moral agency, sensitivity to moral considerations and moral behaviour. If it turns out that current practices of blame and praise are not suited to achieve this effect, then they should be abandoned where possible. This line of thought can also be found in Strawson, when he claims that 'it is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes, in regulating behaviour in ways considered desirable; or to add that when certain of our beliefs about the efficacy of some of these practices turn out to be false, then we may have good reason for dropping or modifying those practices' (2008: 27).

There is some evidence that we should rethink our blaming reactions and practices in particular. For example, Hanna Pickard has stressed the irrational and unhelpful side of our blaming practices and emotions in recent work (2013). The claim that our reactive attitudes can, and possibly should, be subject to change also helps to avoid a certain moral parochialism, which claims that only people who happen to have the same set of moral emotions and reactive attitudes as we do are morally responsible.[6] Not everyone will be happy with the consequentialist reading of Strawson that I am pushing here, following McGeer. Some authors see the ineluctability of our participant reactive attitudes as *the* central feature of his account, rather than the more consequentialist elements that I have been stressing. For the purposes of my discussion, not much hangs on how Strawsonian the account I put forward is. My claim is merely that the consequentialist can adopt a

[6] This issue arises for example in the context of autism and moral responsibility (cf. Richman and Bidshari 2018).

lot of what is attractive about Strawson, whether we thereby remain true to the spirit of Strawson's proposal is a further question. It is likely that my account will still be too one-eyed and consequence-oriented for many Strawsonians.

Desert theorists, too, are likely to object to my account on the basis that, by making responsibility primarily about the reactions that the agent's behaviour merits, I am at most supplying a theory of the justifications of blame and praise. However, responsibility judgements are used to justify (even if only in principle) appropriate reactions, they go beyond the evaluation of actions as bad or good. In my view, this is what sets considerations of responsibility apart from evaluations of actions which proceed by applying a normative ethical theory of moral and immoral action. The free will debate was, and is, partially motivated by the question of what agents need to be like for certain reactions to their behaviour to be justified. Instrumentalist accounts give a distinctive kind of answer to this, which is that agents have to be such as to be able to profit morally from participating in our responsibility practices.

While the proposed account is still clearly instrumentalist, it captures more of the phenomenology of our practices, and because blame- or praiseworthiness does not depend on whether blaming a certain individual will have a positive effect *in that instance*, it also avoids another problematic feature of Smart's account. On Smart's model, an individual instance of ascribing responsibility is justified if we can positively influence the agent's moral conduct by holding them responsible. On the account Vargas and other instrumentalists, myself included, favour, our practices of holding responsible are justified if they over time and on balance tend to make us into more responsible agents. In other words, not every individual ascription of responsibility need have the right effect, it is the aggregate that counts.

> There may be instances where my gratitude or indignation may fail to influence anyone in the proper fashion. Nonetheless, my gratitude (or indignation) 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. (Vargas 2013: 177)

It follows that someone may be an appropriate target of praise and blame even if it is unlikely that a specific *instance* of praise or blame will have the desired effect of contributing to the development of their moral agency. For example, p might not learn from blame by x for having done y because they don't respect x as a moral authority, but this does not change the fact that they are in general susceptible to being influenced by their society's moral responsibility practices and are therefore a morally responsible agent. They can therefore be blameworthy for what they've done even if they do not benefit from blame by a certain person at a certain time.

The refined account is also immune to the criticism mentioned earlier, which is that holding somebody responsible is pure behaviour management of the type we might do with dogs, who aren't morally responsible. We do not expect dogs to understand why what they are doing is wrong, change of behaviour is all we try to effect. The fact that the account aims to develop morally responsible agency and receptivity to moral reasons for action means it is limited to individuals who can in principle be moral agents and understand moral reasons. Dogs do not qualify.

III. IMPLICATIONS

III.1. Being responsible vs being held responsible

We ordinarily believe that whether someone is blameworthy/praiseworthy and whether they should in fact be blamed or praised are two different judgements, and that we can sometimes answer one in the affirmative and the other in the negative. I have already shown that moral influence theories can account for local cases in Section II. We can judge someone to be responsible because they have a certain level of moral understanding and their reasons responsive agency can in general be fostered by holding them responsible, but, nevertheless, it may not make sense for a specific person to hold them responsible in a certain situation.

However, because morally responsible agency is tied to being susceptible to moral influence, people who in principle cannot learn from being held morally responsible are exempt from responsibility on my account. It is sometimes suggested that this is the case with psychopaths, that they are just incapable of 'getting' moral demands (Morse 2008; Shoemaker 2011). There is room for debate on how convincing the empirical case for the moral incapacity of psychopaths is (cf. Godman and Jefferson 2017). However, should it be the case that they really are not amenable to moral influence and are 'morally colour blind', incapable of developing sensitivity to moral considerations, they would to that extent not be morally responsible. That does not mean that they cannot still be legally responsible (cf. Shoemaker 2011), and some of the more blunt instruments in Smart's instrumentalist repertoire may still be applicable. But they would not be *morally* responsible.

Vargas resists this conclusion because he gives separate accounts for what makes agents morally responsible and the justification of the moral responsibility system. 'On the account that I propose, whether the agent is morally responsible for his or her actions is not a function of a 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' (Vargas 2013: 103).

It is true that lack of susceptibility *in a certain situation* is not enough to make the agent non-responsible. But my account entails that general unresponsiveness does exempt. However, exemptions will be very rare, because most humans are receptive to moral responsibility practices, even if this receptiveness is a matter of degree.

Some might object to this on the basis that the person who cannot be influenced still *deserves* blame and therefore is blameworthy, for example because they exhibited ill will. Desert, one might argue, is independent from an agent's capacity to be influenced. But if certain (actual or hypothetical) individuals cannot access and understand certain reasons for action (moral colour-blindness), there is simply no point in trying to engage them in that kind of interaction. If an individual is not a suitable addressee of our responsibility practices, they are not a responsible agent. I am here, once again, in agreement with the Strawsonian, who thinks that for some individuals it may be appropriate to take the objective attitude and shift to a stance of behaviour management. This behaviour management may well involve sanctions in order to act as a deterrent. But it will lack key expectations of moral insight or growth on the side of the person we are trying to influence by our sanction, and we will not have the expectation that these agents acknowledge the wrongness of their action. Clearly, we will want to err on the side of holding individuals responsible rather than merely trying to manage their behaviour, but if certain kinds of interactions and appeals become impossible, then we'll just have to go the 'if you do x, we will do y' route.

This may sound more counterintuitive than it in fact is, because our notions of responsibility are tightly wrapped up with judgements of causal responsibility and evaluations of character. Neither of these are affected. Attributions of causal responsibility and negative evaluations of action and character will remain legitimate. We can still say that psychopaths are depraved, or callous and that their actions were morally wrong, cruel or whatever the applicable thick ethical term is. On some theories of moral responsibility, this is still a notion of moral responsibility, namely attributability responsibility (cf. Shoemaker 2015; Watson 1996). But this is not the kind of responsibility I am interested in. My focus is on accountability responsibility, on what we have a right to expect from others and on the reactions which are appropriate if people do not meet these expectations.

It follows from my claim that morally responsible agency consists in susceptibility to having one's moral agency developed and sustained by moral influence, that those who cannot develop their moral agency through our responsibility practices are not morally responsible.

But there might be scenarios where this seems like the wrong result. A problem case that has been suggested to me is an individual of great moral wisdom who, while remaining imperfect, can develop no further. It would seem that if this individual acts wrongly, they ought not to be blamed on my account,

but this is counterintuitive.[7] This is an interesting case because it provides a putative example of an unpalatable disconnect between having acted wrongly and being blameworthy. However, I would argue that if an agent acts wrongly in avoidable situations, then there is more moral wisdom and self-control to be acquired, and external scaffolding through moral responsibility practices can help with that. In other words, I doubt the psychological plausibility of an individual who cannot be helped in their moral thinking and *acting* by being held responsible. An alternative reading of this case is that these individuals only fail morally in situations which are extremely demanding. If that is the scenario, then blame is indeed unwarranted. But this would be due to the moral demandingness of the task, not due to the fact that the person's moral development is, as it were, complete.

A further implication and potential problem case is that, on my account, talking about the dead being morally responsible is a mistake, because they are past being influenced by our practices. This is tricky case. On one level, my account does imply that we should blame and praise the dead, because in as far as people care about their posthumous reputation, the practice of praising and blaming the dead will help support moral conduct while individuals are still alive.[8] However, at the time that we blame these individuals posthumously, it is too late to influence them.

This means that strictly speaking, the dead are indeed not blame- or praiseworthy. While it is perfectly correct to make the judgement that a certain agent was responsible and was blameworthy during their lifetime, it would be wrong to say that they *still* are. Nevertheless, there are further reasons why people make these kinds of assessments, beyond the useful pre-emptive function of these practices just mentioned. By saying that that person was praiseworthy for what they did, we express our admiration for the action and reinforce our commitment to acting in similar ways. We are also committing ourselves to the claim that the person would be an adequate target for praise and blame, were they alive. This is not to say that we use the term 'blameworthy' as shorthand for 'should have been blamed if alive' in these contexts. Rather, it is part of the self-effacing nature of instrumentalist moral responsibility that outrage and blaming reactions can, as it were, follow people beyond the grave. We may genuinely have certain reactive attitudes towards the dead. But on my account, the correct thing to say regarding dead people is that they acted well or badly and that they were responsible at the time and would have been apt targets of blame. We can also condemn their behaviour, thereby incentivizing the living who care about their posthumous reputation. But literal talk of dead people deserving blame is mistaken on my view.

[7] I thank an anonymous reviewer for pressing me on this case and the case of the blameworthiness of the dead.

[8] Thanks to Jimmy Lenman for pointing this out.

As these examples show, different aspects of the account are linked in such a way that the account of responsible agency does not become untethered from the justification of our responsibility practices. However, arguably, sometimes we *want* our account of responsible agency to be disconnected from that of being a suitable target of our responsibility practices. Sometimes, we want to say that a person is not yet or not fully responsible, but nevertheless a fitting target of our responsibility practices. The most prominent example for this is that of children, but similar issues arise in the context of those suffering from mental disorder (Brandenburg 2018). Traditionally, we endorse both the claim that children are not yet (fully) morally responsible and the claim that they are appropriate targets for our responsibility practices. Being subjected to our moral responsibility practices is precisely how they learn to become responsible agents. Note that this is a different problem from the one I addressed before, which is whether there are times when a generally responsible agent is not a fitting target for a specific instance of being held responsible. Here, the question is whether there can be a disconnect between being a responsible agent and being a fitting target of our *responsibility practices generally*. Vargas has no problem allowing for this, because his theory of responsible agency is separate from his account of what justifies our responsibility practices. However, on my account, allowing for responsible agency and being a fitting target of our responsibility practices to come apart is problematic. And indeed, I will argue that the characterization above that children are non-responsible but suitable targets for being held responsible is inaccurate as it stands.

Let's assume we say that the child is influenceable, as we should, because in the case of children, far more clearly than in other cases, we blame and praise, reward and punish in order to influence their future behaviour. In fact, I think it is no coincidence that Smart's prominent example of the lazy boy involves a child, because the function of behaviour modification and moral education is far more plausible and palatable in the case of children. I point out to little Jane the wrongness of kicking little Jimmy, I tell her off, ask her how she would feel if the same was done to her and quite likely also punish her by withdrawing the half hour of computer games she normally gets in the evening. Does this mean that I take her to be fully blameworthy? Of course not. Responsibility comes in degrees, just as moral agency does.[9]

This is reflected in our judgements, I do not think that Jane is blameworthy in the way I would be blameworthy were I to kick little Jimmy. It is also reflected in our behaviour: we don't treat Jane the same as we would someone older, our blaming behaviour is different. It will be more explanatory and less harsh, at least in the consequences visited on the child. And for children so

[9] This is a fact that has been increasingly recognized in the recent literature; for example, Coates & Swenson (2013) and Nelkin (2016) defend graded notions of moral responsibility and praise- and blameworthiness.

small that they cannot be expected to have a concept of wrongness of action, we do not indulge in moralized blame of punishment at all. Rather, we rely on conditioning, reinforcing desirable behaviour, curtailing and discouraging undesirable behaviour. I do not blame the one year old who hits their playmate or parent at all, I merely take their hand away and say 'no'. It is also worth noting that often, when we hold children responsible, there is an element of play-acting involved, of trying to impress on them the gravity of what they are doing without in fact feeling any of the participant reactive attitudes such as resentment characteristic of full blown holding responsible. Strawson uses small children as an example case of where we take the objective stance and what we do is not a genuine instance of holding responsible. I believe this is right for very young children, but would like to stress the continuity between behaviour management and full-fledged instances of holding responsible. So, rather than saying that in the case of children, we knowingly hold the non-responsible responsible, we should more accurately say that we hold the partially responsible partially responsible. There is now growing recognition of the fact that our practices of holding each other morally responsible lie on a continuum (Brandenburg 2018), and I take the fact that my account allows for this to be a strength. However, a consequentialist account can only do justice to this if it makes the Strawsonian adjustments I have suggested, whereby the kind of interpersonal moral engagement changes as the moral agency of the target of our responsibility practices develops. If all holding responsible were behaviour management, the difference between what we do with small children, bigger children and adults could not be captured. Another important feature of my account that comes out of this is that we cannot say that responsibility is proportional to the susceptibility of being influenced. Otherwise, children would be more, not less responsible than adults. Rather, it is being susceptible to moral considerations and the capacity to be moved and influenced by moral reasons to act or refrain from acting which is relevant to responsibility.

III.2. Worries about instrumentalization and the scope of consequentialism

One of the worries McGeer raises about Vargas' account is that because it has a separate account of what justifies our responsibility practices and of what constitutes a morally responsible agent, there is the danger of a justificatory gap, whereby the general effectiveness of our practices would justify holding someone responsible who does not count as responsible according to the separate account of responsible agency (2015).

Following the discussion above, which illustrates the close link between being susceptible to moral influence and being a responsible agent, one might think this problem is elegantly taken care of in my account. Unfortunately, things are not quite so simple. I have argued that instrumentalist accounts can justify our responsibility practices by looking at the effect on both the

blamers/praisers and those who are blamed and praised. So the target of our practices includes all people who take part in these practices and whose moral agency can be furthered by them. However, this once again raises the worry that the justification for blaming someone might come apart from the agent's responsibility in unacceptable ways. If what justifies blaming and punishing is the overall effect, we might end up with the following situation: assume that the moral community benefits from blaming someone who doesn't in fact meet the criteria for blameworthiness. Could that person not still be justifiably blamed if doing so benefits enough others?

Allowing effects of praise and blame other than those on the person who has done something morally good or bad to justify our practices would seem to reintroduce the common quandary familiar from consequentialist accounts of punishment, that they justify punishing the innocent. A troubling variation of this problem arises on my account—if the person who has done something wrong cannot benefit from being held responsible, they aren't responsible on the criteria for responsibility that I have proposed, but we might still be justified in holding them responsible because of further effects on others. While I cannot fully resolve the issue within this paper, I will sketch what I take to be a plausible response.

It is important to note here that the problem which arises is *not* that of blaming the innocent (so the ones who haven't done anything wrong) but that of blaming the non-responsible (those whose moral agency cannot be affected by being held responsible). The account would not categorize agents as responsible merely in virtue of it being useful to *others* to hold them responsible. Whether a given subject is a responsible agent gets decided by whether they are capable of developing and maintaining responsiveness to moral considerations (reasons responsiveness) through taking part in our moral responsibility practices. Rather, the problem for this account is one of instrumentalization, that it seems as though we might be justified in blaming someone who does not count as responsible if our own or others' agency benefits from doing so. That person would then function as a tool for others' moral development.

Applying this to Smart's example, we may want to hold the stupid boy responsible for his non-achievement because others might benefit from being impressed with the importance of working hard and the negative repercussions of not working hard. Interestingly, this is not a justification Smart considers; rather, he is interested in capturing as much as possible of our judgement that a certain individual is responsible, and that judgement is different from the one that it would be good for others' moral development if we treated them as responsible.[10]

[10] However, as Arneson (2003) points out, this does not mean that, as a Utilitarian, Smart would not have been happy with the idea of holding the lazy boy responsible if the overall effects

At this point, we are faced with four options:

1. We could bite the bullet and concede that blaming and punishing those who don't meet the responsibility criteria is justified if there is enough benefit to others.
2. We could drop the notion that influence on the assessor and the spectator can also justify our responsibility practices.
3. We can try to resolve this question by explicitly positing that being a responsible agent (i.e., being susceptible to moral influence) is a necessary condition for being a fit target for praise, blame, reward and punishment, which cannot be outweighed by any other benefits.
4. We can attempt to show that it does not benefit others' responsible agency to hold the non-responsible responsible.

We should choose options 3 and 4. We should take the benefits for others to be secondary to those for the target of blame, praise, etc. These benefits cannot trump the agent's right not to be blamed when they are not blameworthy (either because they haven't done anything wrong or because they do not meet the responsibility criteria). A standard objection to this is that once one endorses an instrumentalist account of responsibility, one is thereby committed to going the whole hog and applying consequentialist reasoning across the board. It may seem ad hoc to say that consequentialist justifications for moral responsibility justify our practices but are still subject to a fairness constraint, whereby we should not blame the non-responsible. Consequentialists, so the objection goes, cannot help themselves to fairness considerations, as they are all about maximizing the good. Does an instrumentalist account of moral responsibility commit us to consequentialism as an ethical theory? I take this concern to be easily addressed. Being a consequentialist about moral responsibility does not commit one to being a consequentialist across the board, a fact that Vargas stresses repeatedly in *Building Better Beings*. Consequentialist or instrumentalist accounts of moral responsibility address the question: What is the appropriate reaction to morally wrong (or right) action? They do not address the question: What is the correct normative theory of what makes actions right or wrong? Of course, anyone who thinks that consequences should be completely irrelevant to normative theory will struggle with an instrumentalist account of moral responsibility. But the thought that consequences are morally relevant is compatible with further constraints on normative theories, such as, for example, rights.

The worry regarding instrumentalization can be further defused by denying the assumption that we can develop our own moral agency by blaming the non-responsible. Blaming the non-responsible involves moral mistakes regarding

were beneficial. But that would be a further consideration, less tightly linked to the question of our judgement as to whether someone is responsible in the first place.

morally reasons-responsive behaviour and what we can (and should) expect from an agent with a certain set of capacities. Take an example that can frequently be found in the (British) news, where severely autistic, non-verbal teenagers who are also mentally handicapped physically attack their families in situations of stress. These cases are tragic, and restraint and some kind of behaviour management are clearly called for. But it would not develop anyone's moral agency to treat these individuals as fully blameworthy. Instead, it would show lack of compassion as well as lack of understanding of what it takes to be a moral agent. Clearly, whether blaming the non-responsible furthers anybody's moral development is an empirical question, and matters get more tricky in borderline cases, such as the one of psychopaths, where the question whether they are morally responsible is significantly more controversial. Here, I am only sketching the broad outline of an instrumentalist account of moral responsibility and how it should respond to the most important objections against it.

IV. CONCLUSION

I have argued that instrumentalist accounts, suitably revised, can meet the criticism that they describe mere mechanisms of behavioural modification. Behavioural modification is part of what instrumentalist accounts of moral responsibility aim for, but their target is more ambitious, they want to foster moral agency. They can accommodate reactive attitudes in a way Smart's account cannot, though the suitability of our current practices is dependent on whether they actually help us become more reasons-responsive moral agents. However, the account cannot and does not aim to give an account of blameworthiness which is independent of being a suitable object for having one's agency developed through our responsibility practices. This means that moral responsibility is determined by looking forward just as much as it is by looking backwards at what a person has done.

There are two important issues which I have not been able to address in this paper: One is whether the account I propose changes the meaning of the term 'moral responsibility' to the extent that I am, in effect, not talking about what other people are talking about when they use that term. This is an objection frequently levelled against revisionist accounts such as Smart's or, more recently, Vargas' (McCormick 2013; McKenna and Pereboom 2016). It is an important objection, which merits extended treatment. My very short preliminary answer is that most worked out accounts of responsibility involve some amount of revision, as it is unlikely that we have a consistent set of intuitions about moral responsibility. Accordingly, I suspect that the charge of not capturing what we mean by 'moral responsibility' is one which applies not just to card carrying revisionist accounts, and that the more pressing question

is whether the revision is a desirable one that we should endorse. I have argued that it is.

Another issue that merits further discussion is which actions individuals are morally responsible for. To clarify: I have talked about what it takes to be a morally responsible agent, but I have not addressed the question when even morally responsible agents might be excused for doing bad things or shouldn't get credit for behaviour with good outcomes. I have not addressed the effect of coercion, manipulation, stress, etc. on responsibility for action. This is the area where normative ethics and theories of moral responsibility bleed into each other, and it is the bread and butter of many discussions in the literature on moral responsibility. Let me just briefly gesture at the way I think we should go: Our normative theory of choice decides whether and why a certain action is wrong. The extent to which a person is blameworthy when there is an excusing factor such as, for example, coercion will be determined by how difficult it was for them to do the right thing and what we can expect from agents in situations of coercion. If it appears that even a maximally morally sensitive person would have failed at that hurdle, then blame or praise is unjustified and pointless. While more work needs to be done to fully articulate an instrumentalist account of moral responsibility, the purpose of this paper has been to rehabilitate it as a respectable theoretical option by showing how it deals with important objections.[11]

REFERENCES

Arneson, R. (2003) 'The Smart Theory of Moral Responsibility and Desert', in S. Olsaretti (ed.), *Desert and Justice*, 233–58. Oxford: Clarendon Press.

Arpaly, N. (2006) *Merit, Meaning, and Human Bondage: An Essay on Free Will*. Princeton: PUP.

Brandenburg, D. (2018) 'The Nurturing Stance: Making Sense of Responsibility Without Blame', *Pacific Philosophical Quarterly*, 99: 5–22.

Coates, D. J. and Swenson, P. (2013) 'Reasons-Responsiveness and Degrees of Responsibility', *Philos Stud*, 165: 629–45.

Doris, J. M. (2015) 'Doing Without (Arguing About) Desert', *Philos Stud*, 172: 2625–34.

Elzein, N. (2013) 'Basic Desert, Conceptual Revision, and Moral Justification', *Philosophical Explorations*, 16: 212–25.

Godman, M. and Jefferson, A. (2017) 'On Blaming and Punishing Psychopaths', *Criminal Law & Philosophy*, 11: 127–42.

McCormick, K. (2013) 'Anchoring a Revisionist Account of Moral Responsibility', *Journal of Ethics & Social Philosophy*, 7: 1–20.

[11] Earlier versions of this paper were presented at the Gothenburg Responsibility Conference, the Departmental Seminar Series in Sheffield, and the Birkbeck Philosophy Spring Workshop where I received helpful comments. My thanks go to Lucy Campbell, Jan-Hendrik Heinrichs, Michael McKenna, Jimmy Lenman and Philip Robichaud for useful feedback on earlier drafts. I would also like to thank the two anonymous reviewers at *The Philosophical Quarterly* for their helpful comments. I would like to acknowledge the support of the Leverhulme Trust in the writing of this paper and paying open access fees (Grant ID: ECF 2015-493).

McGeer, V. (2014) 'P.F.Strawson's Consequentialism', in D. Shoemaker and N. Tognazzini (eds.), *Oxford Studies in Agency and Responsibility: 'Freedom and Resentment' at 50*, Vol. 2, 64–92. Oxford: OUP.
—— (2015) 'Building a Better Theory of Responsibility', *Philosophical Studies*, 172: 2635–49.
——, and Pettit, P (2015). *The Hard Problem of Responsibility Oxford Studies in Agency and Responsibility*, Vol. 3, 160–88. Oxford: OUP.
McKenna, M. and Pereboom, D. (2016) *Free Will - A Contemporary Introduction*. New York: Routledge.
Morse, S. J. (2008) 'Psychopathy and Criminal Responsibility', *Neuroethics*, 1: 205–12.
Nelkin, D. K. (2016) 'Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness', *Noûs*, 50: 356–78.
Pickard, H. (2013) 'Irrational Blame', *Analysis*, 73: 613–26.
Railton, P. (1984) 'Alienation, Consequentialism, and the Demands of Morality', *Philosophy and Public Affairs*, 13: 134–71.
Richman, K. and Bidshahri, R. (2018) 'Autism, theory of mind, and the reactive attitudes', *Bioethics*, 32: 43–9.
Shoemaker, D. (2015) *Responsibility from the Margins*. Oxford: OUP.
—— (2011) 'Psychopathy, Responsibility, and the Moral/Conventional Distinction', *The Southern Journal of Philosophy*, 49: 99–124.
Shoemaker, D. (2017) 'Response-Dependent Responsibility; or, A Funny Thing Happened on the Way to Blame', *Philosophical Review*, 126: 481–527.
Smart, J. J. C. (1961) 'Free-Will, Praise and Blame', *Mind*, 70: 291–306.
Strawson, P. F. (1974/2008) 'Freedom and Resentment', in P. F. Strawson (ed.), *Freedom and Resentment and Other Essays*, 1–28. Abingdon: Routledge.
Vargas, M. (2008) 'Moral Influence, Moral Responsibility', in N. Trakakis and D. Cohen (eds.), *Essays on Free Will and Moral Responsibility*, 90–122. Newcastle: Cambridge Scholars Press.
—— (2013) *Building Better Beings: A Theory of Moral Responsibility*. Oxford: OUP.
Wallace, R. J. (1994) *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
Watson, G. (1996) 'Two Faces of Responsibility', *Philosophical Topics*, 24: 227–48.

University of Birmingham, UK