GET SMART: OUTCOMES, INFLUENCE, AND RESPONSIBILITY

Abstract: Once relegated to the margins of the responsibility debate, moral influence theories have recently been rehabilitated. This paper offers a moral influence theory with two parts: a theory of responsibility as influenceability and an act consequentialist justification of blame. I defend this account against six concerns commonly raised both by opponents and by advocates of similar views. Some concerns target act consequentialism, claiming that it 1) permits blaming innocents; 2) permits coercion, manipulation, and other objectionable forms of influence; and 3) fails to capture intuitions about desert. Other concerns target responsibility as influenceability, claiming that influenceability accounts are 4) unsophisticated, 5) make ascriptions of responsibility dependent on assessments of permissible blame, and 6) have various counterintuitive implications.

For decades moral influence theories of responsibility were rejected, dismissed, and ignored. They were maligned by eager Strawsonians and marginalized as many moral theorists turned decisively away from utilitarianism. Recently, though, these theories have been rehabilitated and are now a going concern. But this revival displays a curious pattern: fresh formulations of classic claims typically make their case by distancing themselves from purportedly problematic predecessors. In particular, they eschew the consequentialism they associate with classic theories. Distilling the various challenges to their essence, it is this: *these theories overreach*. (Perhaps, being one-eyed, they lack the depth perception necessary to grasp the truth.) They try to do too much with too little (Vargas 2013, 171). They stretch a simple theory over the knotty complexities of our moral world, with uncomfortable results. My aim is to

resist this critique and to argue that classic theories (e.g., Schlick 1939; Smart 1961) have the resources to respond to the challenges raised by their opponents (e.g., Bennett 1980; Wallace 1994; Scanlon 1998) and echoed by other instrumentalists (Vargas 2013; McGeer 2015; Barrett 2020).

I take moral influence theories of responsibility have two parts: an influence theory explains responsibility in terms of influenceability, and an instrumentalist theory justifies blame in terms of its consequences. The aim of this paper is to articulate and defend one such theory. Section 1 introduces the concepts necessary to make this evaluation. The next two sections introduce the two parts of the theory, which I call *Outcome Influenceability:* section 2 defends act consequentialism as the relevant instrumentalist theory and considers three problems for it as an account of how blame is justified; and section 3 defends an account of moral responsibility as influenceability and considers three challenges to its adequacy. Together these two sections describe the structure of moral influence theories so that their commitments can be identified as precisely as possible and compared with analogous commitments in traditional responsibility theories. Having done so, they demonstrate that Smart-style theories are plausible contenders and warrant further support and refinement. Section 4 concludes.

1. Preliminaries

The motivating idea behind influence theories is that, in order for an agent to be responsible for their misconduct, it must be possible to prevent or deter such behavior. In this respect, influence theories resemble other responsibility theories according to which an agent must, for example, be responsive to the reasons for and against that act, where responsiveness is a kind of influenceability. Of course, influence theories are importantly different from other

popular theories. Most controversially, they allow a wider range of influenceability to bear on responsibility. In particular, they are more permissive with respect to *when* and *how* the agent can be influenced and still count as responsible. I will say more about this permissiveness below (§3.2).

Instrumentalism about responsibility is the view that whether, when, and how we blame and praise ourselves and others is justified by their good outcomes.[1] Instrumentalism is not the same as consequentialism; all consequentialist views are instrumentalist, but not all instrumentalist views are consequentialist. One might hold that blame is justified by its good consequences, but admit restrictions on how we pursue good consequences (Tadros 2011, 13). Most instrumentalist accounts of responsibility are explicitly revisionist (Smart 1961; Arneson 2003; Vargas 2013). Revisionism says that a plausible account of moral responsibility will be at odds with common sense because many of our ordinary convictions about what it means for a person to be responsible for their conduct are problematic and must be abandoned (Vargas 2013, 2).

The distinction between influence theories and instrumentalist theories tracks the distinction between what Victoria McGeer calls the presupposition question and the justification question, respectively (2014, 67-68). It also roughly tracks the distinction between conditions on *being* responsible and conditions on the appropriateness of *holding* an agent responsible (Wallace 1994, 91).

To hold responsible is to react to and treat a person in particular ways in response to their behavior. A practice of holding responsible consists in the pattern of reactions and treatment received by those we hold responsible and the expectations this pattern generates in people. Blame is one way to hold responsible, though not the only way. However, while there may be

ways of holding responsible that function like blame, without rising to the level of blame (Pickard 2011), this paper will focus on blame and the question of when, if ever, blame is justified.

Blame is a particular kind of response to an agent's wrongdoing. As with punishment, some responses to wrongdoing count as blame while others don't. The criminal legal system may punish a person who, unemployed and struggling to support their children, robs their neighbor's house. They may go to prison and have to compensate their neighbor for the stolen property. But a just legal system may respond in other ways, too: they may receive therapy while in prison and a housing subsidy upon release. The latter responses are not part of their punishment. Similarly, one may blame a neighbor who, frustrated by daily protests and the increasingly bitter rhetoric about them, appeals to a racist stereotype during a political conversation. One may express indignation and disgust or give them the cold shoulder for the rest of the evening. But one might respond in other ways, too: one might try to explain why the stereotype is mistaken, hurtful, and dangerously persistent. The latter responses are not blame.

Whether to blame is a moral issue because blame is (usually) a harm and is (usually, at least somewhat) voluntary (McKenna 2012; Shoemaker 2020). It hurts when one is blamed and one can often refrain from blaming or cease to blame in many circumstances even if one's initial blame response is involuntary.[2] We have more control over directed blame, which has a public, expressive, or communicative dimension, than we do over private blame, which is an internal, attitudinal response. But we have some control, direct and indirect, over both. I will focus on directed blame.

2. Maximizing Act Consequentialism

Instrumentalist theories of responsibility justify blame by reference to its consequences. However, few are act consequentialist. Indeed, instrumentalists often include, among their *bona fides*, a litany of act consequentialism's supposed inadequacies. An evaluation of Outcome Influenceability, of which act consequentialism is one part, must therefore assess these concerns.

I don't pretend to offer a complete defense of act consequentialism. I ignore some of its opponents' main worries—e.g., about demandingness and the separateness of persons—and focus instead on objections to its suitability for the responsibility debate. I want to assess whether act consequentialism is fit for use in a theory of moral responsibility.

> *Maximizing Act Consequentialism (MAC):* An act is morally permissible just in case it promotes as much good, impartially considered, as any other available act.[3]

There are a few things to note about MAC. First, the theory applies to all human behavior. It takes *all* good outcomes for *all* sentient beings as bearing on the question of whether to blame. Thus, MAC directly answers the justification question: Finn should hold Jake responsible for his misconduct just in case and in the manner that promotes as much good as any other available act. Second, as a moral theory, MAC is neutral with respect to value theory—i.e., what it is in virtue of which a state of affairs is morally good or good for some individual—and with respect to moral psychology—i.e., which facts, mechanisms, theories best describe and explain human (moral) psychology, including (moral) cognition, (moral) emotion, and (moral) motivation.

2.1. The Innocence Problem

A common charge against instrumentalist theories is that they permit knowingly blaming the innocent (Wallace 1994, 57; Scanlon 1998, 267; Shoemaker 2020, 225-226).[4] In response, many instrumentalists have embraced versions of instrumentalism that avoid this implication.

Some have argued, following Rawls, that the *practice* of blame, rather than individual acts of blaming, is justified by its consequences (Vargas 2013, 194; Barrett 2020, 10).

Act consequentialism can also avoid the innocence problem. It does not require that the decision (e.g., to blame) be directly motivated by an interest in maximizing the good. Simon might justifiably believe that the best possible world is one with friendships in it, but also realize that he is most likely to achieve his aim not by directly trying to create the best friendships, but by pursuing friendship for its own sake and indirectly bringing about the best world. This procedure would be acceptable to a sophisticated act consequentialist, even if it occasionally yields suboptimal decisions, so long as the world is better overall than if one lacked these values and motivations (Railton 1984). McGeer (2014) has argued that an act consequentialist can justify a fairly traditional responsibility practice, that eschews blaming the innocent, in the same way.

Another response is to point out that blame is necessarily a *response* to a person's behavior—just as punishment is an imposition on an offender *for an offense* (Duff 2001, 96). If Marceline is upset about Finn's behavior and intentionally treats Jake badly despite knowing that he is entirely innocent, she is not blaming Jake. She's just treating Jake badly. (She might unconsciously blame him, despite believing that he's innocent, but that's not what I'm describing.)

Of course, MAC *does* permit harming known innocents, including by treating them *as if* blaming them, in order to promote the greatest good. Marceline might try to trick others into believing Jake was responsible by acting in ways that signal blame. But if she knows that he hasn't done anything wrong, then she is not blaming him, however it might seem to others. That it permits imposing harm on some innocents for the sake of other innocents is not a decisive

objection to MAC. Doing so appears to be widely accepted—or at least widely done. States impose burdens on some in order to promote greater benefits for others via preemptive war, seatbelt laws, immigration policy, and vaccination requirements. So do individuals. Essential workers accept long hours and endanger their families in order to help those in urgent or severe need; and generous people give lots of time and money to needy strangers thereby depriving their friends and loved ones of attention and assistance they might reasonably have expected. According to MAC, such acts are permissible because the claim of one innocent person to avoid the harm resulting from some act (e.g., *as if* blaming) does not outweigh another innocent person's claim to avoid the similar harm resulting from not acting (e.g., not *as if* blaming).

Those who remain uneasy about harming known innocents must acknowledge that MAC is not alone in justifying such harms. Any moral theory that accepts that consequences are intrinsically morally relevant (i.e., bear directly on moral permissibility) and that constraints on doing harm are not absolute, must accept that, at least in principle, harming the innocent can be permissible. The real disagreement is not about whether harming innocents is ever justified, but rather how much innocence counts for in justifying a harm (Ross 1930). However, this disagreement about the moral significance of innocence does not challenge the claim that MAC is fit for use in a theory of moral responsibility. It's fit for use, despite being a controversial moral theory.

2.2. The Influence Problem

Another common charge is that instrumentalist theories permit objectionable methods of influencing people—e.g., deception, manipulation, and coercion. In the context of the responsibility debate, MAC in particular seems to permit unjustifiable forms of blame. P.F.

Strawson, for example, can be read as worrying that, insofar as they advocate taking the objective stance toward offenders, one-eyed utilitarians like Schlick and Smart advocate 'managing 'or 'treating 'them rather than engaging them rationally from the participant stance (1962, 163; Shoemaker 2020, 225)—think Alex in *A Clockwork Orange.*

Defenders of MAC can appeal to the usual responses. Sophisticated act consequentialism might ensure that only appropriate methods of influence are deployed. Contingent facts about their expected value might rule out coercive and manipulative influence, either because they are clearly suboptimal or because uncertainty supports a cautious attitude toward their deployment. After all, it's almost always easier to ask a person to be more considerate than to try to manipulate or coerce them into acting better. However, these responses arguably duck the issue. The force of the influence problem comes from the idea that respect for persons requires that we attempt to influence them rationally (e.g., via persuasion), rather than through physical or merely psychological means (e.g., via quarantine or nudges). However, the challenge has less bite when it comes to blame, which is usually understood as a form of rational influence. It is either a form of moral address or communication (McKenna 2012); or, like punishment, provides prudential reasons to behave appropriately (Duff 2001, 86-87). Insofar as we're interested in whether and when MAC permits blame, the problems with these other forms of influence are immaterial.

However, blame often involves other kinds of influence, too. Moral anger may express grievances and address them to a wrongdoer, but it may also induce non-rational psychological responses (e.g., fear, embarrassment, hostility) and confront its target physically or bodily (e.g., via the volume or tone of voice, typical gestures and postures of anger, and the bodily reactions they cause). Indeed, it seems exceedingly difficult not to exert all three kinds of influence at once. Nor is blame the only phenomenon, or even the only form of communication, about which

this is true. Public speaking and flirting both take advantage of all of these means—proximity, eye contact, gestures, mirroring—and can be deployed more or less consciously and purposefully. Insofar as we permit blame, we permit these non-rational features of blame.

Nonetheless, the influence problem persists. Act consequentialism must admit the possibility of cases where coercion and manipulation are permissible. Moreover, while non-consequentialist theories permit rational blame that exerts physical and merely psychological influence as an unavoidable side effect—a doctrine of double effect for blame—MAC denies the relevance of this distinction. These are real worries and central to ongoing debates in normative ethics. Again, though, they do not show that MAC is unfit for use in a theory of moral responsibility.

2.3. The Desert Problem

Another charge against instrumentalist theories, including MAC, is that they cannot capture the common intuition that blame is justified when and because it is deserved (Vargas 2013, 250). If MAC is true, then it is not the case that being responsible for a bad act directly contributes to the permissibility of blame. For some this makes MAC a non-starter as a justification of blame.

We should distinguish this challenge from the worry that MAC cannot make sense of desert or distinguish the deserving from the undeserving. If the desert base of blame—that in virtue of which blame is deserved—is responsible wrongdoing, then any account of responsible wrongdoing can *explain* when blame is deserved. While MAC is not such an account, it can be attached to one, as we'll see in the next section. The problem for MAC is how desert can be held to *justify* blame.

Responding to this problem requires distinguishing two justification questions. The first concerns desert directly. When does a person deserve blame? The answer is that blame is deserved only when one is responsible for some wrongdoing. The other question concerns moral permissibility generally. When is it overall morally permissible to blame an agent for their responsible wrongdoing? Answering this question requires that we consider (other) moral reasons for and against blame, including the consequences of doing so. Desert is usually taken to be a moral reason to blame, and a pro tanto reason rather than a decisive one.[5] The fact that blame is deserved does not determine whether blame is morally permissible, much less whether it's appropriate all things considered. The relationship between desert reasons, outcome reasons, and any other moral reasons to blame is an open question in normative ethics. Some accounts of the moral significance of desert are incompatible with MAC, others are not. However, even the latter accounts can recognize the fact of desert and allow that the deserving are fitting targets of blame. Whether act consequentialist accounts of the permissibility of blame can find a plausible role for desert is not clear, but MAC remains fit for use in a theory of responsibility.

…

Each of the three problems discussed above assumes that being responsible is one factor that determines whether blame is justified and that MAC denies this. I've tried to show that, while it does deny the intrinsic moral significance of innocence and of intent, MAC, as a moral theory, can capture the significance of responsibility. But *whether* and *how* MAC takes responsibility to be significant depends on what responsibility is! In the next section, I'll show that there is an account of responsibility that is plausible and that fits with an act consequentialist justification of blame.

3. Influenceability

The previous section addressed challenges to instrumentalist justifications of blame and to MAC in particular. This section addresses challenges to an account of responsibility as influenceability. The fact of such challenges might seem odd. If MAC gives a complete account of whether, when, and how much to blame, why does it need a responsibility theory? Suppose Marceline is a hedonistic act utilitarian who already knows the total utility of blaming Jake for stealing. How is Jake's responsibility status relevant to her decision? Strictly speaking it might not be. But one can only know that if one knows whether Jake is influenceable via blame. This insight is the foundation of a responsibility theory that tries to capture the moral significance of being responsible in a way compatible with MAC (or any other instrumentalist account).

The fact that Jake is responsible may be indirectly relevant to deciding whether to blame because responsibility, like friendship, is something that agents have reason to value in its own right when so valuing will lead them to do the most good (McGeer 2014). Or perhaps the fact that Jake is responsible and Finn is not makes it better to blame Jake, all else being equal. Or it may just be that a responsibility theory is useful insofar as it tells us to look for agents who can be influenced by blame. A responsibility theory is a useful guide to implementing an instrumentalist account of justification.

As I see it, responsibility theories supplement moral theories in the same way that theories of moral status supplement moral theories. Moral theories tell us how to treat beings with moral status. Accounts of moral status explain what it means to have this status (conceptual insight) in terms of having particular properties (e.g., sentience); they thereby allow us to identify beings that have this status (empirical insight). Thus, an account of how we should treat different beings requires a defense of the claim that some property (e.g., sentience) grounds

moral status (i.e., is morally significant) *and* an account what having that property consists in (i.e., what sentience is).

Likewise, an adequate instrumentalist account of moral responsibility requires a complementary account of responsibility. In this section, I defend one such account against some common objections. If this defense is successful, then we have reason to think that Outcome Influenceability, a version of the Smart theory of responsibility that combines act consequentialism with influenceability, is plausible.

> *Influenceability*:* X is morally responsible for an act A iff X is i) an agent, that is ii) causally responsible for A, and iii) influenceable, iv) via their own agency (e.g., via their responsiveness to moral considerations).[6]

Notice the scope of this claim. It's a claim about agents. Asteroids and house plants can be influenced, but cannot be responsible. It's also a claim about agents who can be influenced through their own agency. Some animals and algorithms may be agents, but are not influenceable through their own agency.

Finally, before considering challenges to this claim, let me note a few features of the view. First, being influenceable in one way (e.g., via a fine), doesn't entail being influenceable in another way (e.g., via blame). Second, the claim says nothing about which form of influence is appropriate in cases where multiple forms would be effective (e.g., a fine vs. blame). Third, its structure allows that one can be a responsible agent without being responsible for a particular act—e.g., it allows that a responsible agent could be excused of wrongdoing if physical restraint was the only way to prevent them from committing their offense. Fourth, one can be more or less responsible because more or less influenceable; and circumstances can diminish or enhance the efficacy of holding responsible.

3.1. The Sophistication Problem

The most common challenge to influence theories of responsibility is that they are unsophisticated. Strawson and his intellectual descendants have raised this objection against classic influence theories, but contemporary instrumentalists deploy it to distinguish their updated accounts from supposedly deficient predecessors.

The sophistication problem takes two main forms. One charge is that influence theories assume an unsophisticated conception of human moral and empirical psychology that does not adequately describe the phenomenon. Manuel Vargas claims that classic influence theories offer a detached or therapeutic approach that fails to capture how we actually blame (2013, 169; see also Bennett 1980; McGeer 2015; and Barrett 2020).[7] Another charge is that they advocate unsophisticated means of responding to wrongdoers. R. Jay Wallace claims that influence theories understand our blaming practice as an 'economy of threats' that ignores the attitudinal aspect of blame and fails to explain the force and quality of blame (1994, 56; see also Bennett 1980; Scanlon 1998; and even Jefferson 2019, 558).[8]

Influenceability* can avoid both of these worries. With respect to the first problem, Richard Arneson (2003) has pointed out that influence theories are neutral with respect to moral and empirical psychology.[9] They can accept whichever theories are most plausible, sophisticated or not. Moreover, Outcome Influenceability (i.e., Influenceability* plus MAC) has especially good reason to accept subtle well-supported psychological theories. Sound theories are necessary to reliably produce good outcomes, so influence theories should be expected to accept those theories insofar as they are instrumentalist. Similarly, we should expect them to accept whichever accounts of moral judgment, emotion, and motivation best describe our responsibility

practices and best predict effective uptake of blame. In short, proponents of Influenceability* can and should accept the best available account of when, why, and how blame works. And if this story is different for different forms of blame—moral anger, anticipated guilt, internet shaming, social exclusion—Influenceability* can accept these nuances too.

What about the charge that classic influence theories advocate unsophisticated methods of influencing people, a stick-and-carrot economy of threats? This is simply not the case. Again, influence theories do not advocate *any* method of influence unless combined with a particular instrumentalist account of what justifies blame.[10] And, as I discussed above (§2.2), instrumentalists can accept whichever methods of influence most effectively promote good outcomes. The proximal target of influence need not be agents' behavior; beliefs, attitudes, emotions, dispositions, and capacities are all relevant targets because they all contribute to promoting the overall good. If the best way to achieve the relevant outcomes is to feed people carrots and hit them with sticks, then some instrumentalist influence theories will advocate doing so. If it is more effective to let our responsibility practices operate exactly as they presently do, then those theories will advocate that instead. Whatever we discover—in psychology, neuroscience, cognitive science, evolutionary biology, cultural anthropology, or economics; about persuasive technology, the ethics of risk, or the plasticity of human nature—Influenceability* can acknowledge. Influence theories needn't understand blame solely in terms of carrots and sticks any more than economists must understand consumer behavior as determined solely by price curves.

Consider a particular worry that appeals to both forms of the sophistication problem. Strawson and others suggest that influence theories view blame as a conscious strategy for 'managing' problematic individuals. But this is not the case. Influenceability* allows that blame

influences its object even if one is not conscious that it is doing so, and that even conscious blame need not be pursued with a particular outcome in mind. Moreover, while MAC permits influencing offenders in ways that bypass their agency, it does not advocate managing people rather than engaging them in the sophisticated and rational ways that blame often can. Finally, it does suggest some forms of management—e.g., that we consciously and purposefully cultivate our dispositions to blame and monitor our own and our society's blaming practices in order to blame optimally—but this is to promote a *more* sophisticated approach to blame and is a reasonable suggestion given our susceptibility to epistemic and moral mistakes.

Thus, while the sophistication problem may be the most frequent challenge to influence theories, it seriously misunderstands and underestimates views like Influenceability*.

3.2. The Independence Problem

Another concern is that influence theories depend on a complementary theory of what justifies blame and that instrumentalists construct their influence theories according to the dictates of their preferred moral theory. This is a less common objection, but it's important to address because it perpetuates a misunderstanding of the relationship between responsibility theories and moral theories. Vargas is particularly keen to avoid this kind of dependence, identifying it as one of the main weaknesses of classic moral influence theories (2013, 126-130).[11] However, while there are different ways to understand the worry, none of them undermines the case for Influenceability*.

If the demand is that a responsibility theory be able to be integrated with different moral theories, then Influenceability* meets this demand. It's consistent with any moral theory and it fits particularly well with moral theories that give greater weight to outcomes in determining the

permissibility of blame, including consequentialism and forms of moderate deontology. Moreover, even if Influenceability* is in tension with some moral theories, so are other responsibility theories, as Vargas notes with respect to libertarianism (2013, 70-72).

What about the worry that classic influence theories allow their moral commitments to dictate the structure of their responsibility theory (Vargas 2013, 129)? Many take J.J.C. Smart's responsibility theory to be determined by his act utilitarianism, though Smart never mentions utilitarianism in his discussion of responsibility (Arneson 2003, 244-245). Even if the worry were justified, it does not condemn accounts like Influenceability* because the same charge applies to many of its competitors. If we assume that being responsible bears on the justification of blame, then, however we formulate our responsibility theory, we must be able to make sense of that assumption in light of the nature of responsibility. For example, if one claims that basic desert is sufficient to justify blame, one's responsibility theory must explain—e.g., in terms of control—how responsibility is relevant to that claim. To use Vargas' labels, there may not be a 'limited ethics 'approach to responsibility theorizing. And if not—if we must choose between a responsibility theory dictated by one's moral theory (an 'ethics partisan' view) and a moral theory dictated by one's responsibility theory (a 'responsibility partisan' view)—it's not obvious that one kind of dependence is preferable.

McGeer (2015) offers another objection to Vargas' 'independence constraint'. Suppose that Simon acted badly yesterday and that Finn responds to his behavior today with anger. What grounds Finn's implicit ascription of responsibility? Traditional theories explain Simon's responsibility by reference to facts about him *yesterday* (e.g., that he was reasons-responsive). But McGeer claims that these facts cannot explain why we should blame Simon for his failure rather than, say, console him about it. Why isn't it just unfortunate that, despite having the

relevant capacity, Simon failed in this instance to respond to the relevant moral considerations? An influence theory can tell us. It makes sense to blame Simon if (and because) blame will influence him (and others) to act better in the future. It is facts about Simon *today* that explain his responsibility.[12] One might object that the description of Simon yesterday is not a description of a responsible agent. McGeer responds that this is precisely the mistake. Responsibility ascriptions should be understood not as *descriptions* of agents at the time of *action*, but rather as *exhortations* of agents at the time of *reaction*. This is what makes her theory a dependence theory (specifically an 'ethics partisan' theory). Simon is responsible for his misconduct in virtue of the fact that a blaming exhortation will make him more sensitive to the relevant moral reasons in the future (2015, 2647).

McGeer makes her case against Vargas' independence constraint by showing that replacing one condition (reasons-responsiveness at the time of action) with another (influenceability at the time of blame) yields a more plausible responsibility theory. We can reach a similar conclusion about dependence—and further explicate Influenceability*—by considering how its structure might or might not depend on MAC. First, consider how Influenceability* diverges from the presumed dictates of MAC. As a universal moral theory, MAC is concerned with *all* forms of influenceability and *all* opportunities to influence. It should therefore prefer a responsibility theory according to which *X is responsible iff X is influenceable (at any time, in any way)*. But this isn't the best influence theory because it's just not what people mean when they refer to responsibility; there are many ways of influencing people that just don't fall under our concept of holding responsible—e.g., exploiting a cognitive bias. Nor does this claim violate MAC; those other forms of influence still matter for our decisions, they're just not part of the responsibility story.

An adequate influence theory must restrict the forms of influenceability that determine responsibility and must do so in a non-arbitrary way. Influenceability* picks out *agents* who can be influenced *through their own agency*. These are not arbitrary restrictions. If nothing else, they're restrictions that most responsibility theories accept. Responsibility theorists use intuitions about the boundaries of the concept to restrict the kinds of influenceability that count as responsibility (e.g., reasons-responsiveness). To avoid arbitrariness we identify the properties that explain those intuitions; according to many theories, including Influenceability*, those properties are agential control and/or self-determination.

Nevertheless, Influenceability* may depend on MAC, in an attenuated way. We can see this by considering how different instrumentalists might prefer different influence theories depending on their moral commitments. For example, consider the restriction that, in order to be morally responsible for A, X must be influenceable *at the time they do A.* If we assume that being morally responsible bears on the justification of blame, MAC would view this restriction as arbitrary and unjustified. If blame achieves the same outcome—e.g., enhancing Simon's moral considerations responsiveness and deterring misconduct through his agency—why should it matter that he wasn't influenceable at the time of action? The traditionalist answer is that, as described, Simon doesn't *deserve* blame. However, if desert status is explained by responsibility status—Simon deserves blame (or not) because he's responsible for some misconduct (or not)—then this answer is question begging. It says that, in order to be morally responsible, Simon must deserve blame, but he doesn't because he's not morally responsible. So desert status must be explaining responsibility status—Simon is responsible for some misconduct (or not) because he deserves blame (or not). But this version of the traditionalist answer acknowledges that the structure of a responsibility theory is dictated by a moral theory, namely, the view that the

permissibility of blame depends on desert. Whether or not this is plausible, it directly undermines the challenge posed by the independence problem. Influenceability* is an admittedly controversial and revisionist account, but it is not the only responsibility theory that relies on an (often implicit) complementary moral theory.

A responsibility theory is an account of which properties determine responsibility and why those particular properties are relevant. Influenceability* says that a particular kind of influenceability is relevant because it's a feature of the offender that bears on why blame is justified. Influenceability bears on the justification of blame because, according to some moral theories, blame is justified by its good outcomes and blaming the influenceable usually produces the best outcomes. By contrast, Reasons-Responsiveness says that responsiveness to reasons is relevant because it's a feature of the offender that bears on why blame is justified. Responsiveness to reasons bears on the justification of blame because, according to some moral theories, blame is justified by desert and reasons-responsive wrongdoing is sufficient for deserving blame. Thus, with respect to the independence problem, we can acknowledge that Influenceability* does depend in part on MAC, but also see that this is much less objectionable than it might have seemed.

3.3. The Counterexample Problem

In light of various concerns—those just discussed, but also others—influence theories are claimed to have counterintuitive implications, attributing responsibility to those who are not (false positives) and failing to attribute responsibility to those who are (false negatives). Revisionist accounts inevitably diverge from common opinion in some ways—the failure of traditional theories to accommodate common assessments is precisely why revisionism is on the

table (Arneson 2003, 249)—but revisionists must still respond to hard cases. We can evaluate Influenceability* by considering how it handles supposed false positives and false negatives.

Influence theories can respond to counterexamples in different ways: they can a) accept the counterexample and alter the theory to avoid it, or they can reject it either by b) showing that it doesn't follow from the theory, or c) by arguing that the apparently counterintuitive result is more plausible that it seems.

In response to some counterexamples, influence theorists have used the first strategy to rehabilitate classic accounts. McGeer and Pettit seem to accept the requirement that an agent have been, at the time of action, self-directed and responsive to reasons (2015, 185-186). Similarly, I have argued that, in order to be morally responsible, one must be influenceable in particular ways—e.g., that an agent must be influenceable through their own agency (§3.2). Restricted in this way, Influenceability* is less likely to yield false negatives and false positives. For example, agents influenceable only through manipulation or physical restraint will not count as responsible, nor will those animals who are only influenceable via agency-bypassing forms of conditioning.

For other counterexamples the second strategy seems more promising. Arneson argues that influence theories can affirm sophisticated accounts of moral and empirical psychology (2003, 240). (Others make similar arguments, but understand this move as a revision of an otherwise inadequate theory [McGeer 2014; Barrett 2020]). Recognizing that responsibility is scalar can also defuse common counterexamples to influence theories. Like any adequate theory, Influenceability* allows that immature and impaired agents may be partially responsible, responsible only in certain domains, or responsible only at certain times. It is both accurate and appropriate to ascribe responsibility to some children and people with mental illness rather than

'treating them like children'. A young child may be blameworthy for snatching her toy from a friend, but not for pretending to run away, hiding all day from her parents, and causing them to worry. She knows that snatching and refusing to share is mean, but lacks the foresight to recognize that pretending to run away might distress her parents. Likewise, a person whose mental illness is causing her to worry that she is being stalked by a dangerous demon may engage in magical protective rituals that demonstrate a responsibility undermining impairment; however she may still have the ability to treat others with care and respect, and can therefore be responsible for, say, treating a cashier unkindly.[13] As a scalar theory, Influenceability* can also help determine whether, when, and to what degree such agents are responsible enough to blame, namely, by determining the level at which doing so produces helpful results.

Finally, influence theorists can pursue the third strategy by rejecting dubious orthodoxies and showing that supposedly questionable responsibility attributions are actually plausible—recasting a bug in the theory as a feature of it. One dubious orthodoxy is the assumption that ascriptions of responsibility are just descriptive reports of a capacity rather than exhortations to display it (McGeer and Pettit 2015, 178). Likewise, Anneli Jefferson defends a requirement that an agent is capable of becoming, developing and maintaining moral considerations responsiveness rather than being responsive (2019, 569-570). Some characterize this strategy as simply accepting the surprising implication and refer to it as 'outsmarting'—as in out-Smarting—the counterexample. Suppose that Simon has acted badly, but that the offense itself has completely convinced everyone, including him, to resist and condemn such behavior in the future (Gallagher 1978, 19). Neither Simon nor anyone else is influenceable via blame because no more influence is needed. Influenceability* has the admittedly surprising implication that Simon is not responsible. But is this really so implausible? Such a person has either been scared

straight or reformed, in which case they are arguably no longer blameworthy. (Some might call this 'biting the bullet,' but to me that implies accepting the implication without defending its plausibility.)

4. Conclusion

I have articulated and defended Outcome Influenceability as a plausible theory of moral responsibility capable of addressing the most common and compelling challenges to classic and contemporary influence theories, including those raised by other instrumentalists. Admittedly, my defense has been programmatic. There is more to be said for and against both Outcome Influenceability and other instrumentalist influence theories. However, I hope to have laid out the structure of such views and of their most frequent challenges in such a way that more precise critiques, revisions, and defenses can be developed.[14]

REFERENCES

Arneson, R.J. 2003. "The Smart Theory of Moral Responsibility and Desert" in Olsaretti, ed. (2003, 233-258).

Barrett, J. 2020. "Optimism about Moral Responsibility." Philosophical Imprints.

Bennett, J. 1980. "Accountability" in Van Straaten ed. (1980, 14-47)

Coates, D.J. and N.A. Tognazzini, eds. 2019. Oxford Studies in Agency and Responsibility vol. 5. Oxford: Oxford University Press.

Duff, R.A. 2001. Punishment, Communication, and Community. Oxford: Oxford University Press.

Eggleston, B. and D. Miller, eds. 2014. The Cambridge Companion to Utilitarianism. Cambridge: Cambridge University Press.

Eggleston, B. 2014. “Act Utilitarianism” in Eggleston and Miller, eds. (2014, 125-145).

Gallagher, N. 1978. “Utilitarian Blame: Retrospect and Prospects.” Journal of Value Inquiry 12.1, 13-23.

Hart, H.L.A. 1968. “Legal Responsibility and Excuses” in Punishment and Responsibility: Essays in the Philosophy of Law. Oxford: Oxford UP, 2008: 28-53.

Hieronymi, P. 2019. “I’ll Bet You Think This Blame is about You” in Coates and Tognazzini, eds. (2019, 60-87).

Jefferson, A. 2019. “Instrumentalism About Moral Responsibility Revisited.” Philosophical Quarterly 69, 555-573.

McCloskey, H.J. 1965. “A Non-Utilitarian Approach to Punishment.” Inquiry 8, 249-263.

McGeer, V. 2014. “P.F. Strawson’s Consequentialism” in Shoemaker and Tognazzini, eds. (2014, 64-92).

McGeer, V. 2015. “Building a Better Theory of Responsibility.” Philosophical Studies 172, 2635-2649.

McGeer, V. and P. Pettit. 2015. “The Hard Problem of Responsibility” in Shoemaker Tognazzini, eds. (2015, 160-188).

McKenna, M. 2012. Conversation and Responsibility. Oxford: Oxford University Press.

Miller, D.E. 2014. “Rule Utilitarianism” in Eggleston and Miller, eds. (2014, 146-165).

Miller, D.E. 2014. “Reactive Attitudes and the Hare-Williams Debate: Towards a New Consequentialist Moral Psychology.” The Philosophical Quarterly 64.254, 39-59.

Olsaretti, S. ed. 2003. Justice and Desert. Oxford: Clarendon Press.

Pickard, H. 2011. “Responsibility without Blame: Empathy and the Effective Treatment of Personality Disorder.” Philosophy, Psychiatry, and Psychology 18.3, 209-223.

Railton, P. 1984. “Alienation, Consequentialism, and the Demands of Morality.” Philosophy and Public Affairs 13.2, 134-171.

Ross, W.D. 1930. The Right and the Good. Oxford: Clarendon Press.

Scanlon, T.M. 1998. What We Owe to Each Other. Cambridge, MA: Harvard University Press.

Schlick, M. 1939. Problems of Ethics. New York: Prentice-Hall.

Shoemaker, D. 2020. “Responsibility: The State of the Question.” Southern Journal of Philosophy 58.2, 205-237.

Shoemaker, D. and N.A. Tognazzini, eds. 2014. Oxford Studies in Agency and Responsibility vol. 2. Oxford: Oxford University Press.

———. 2015. Oxford Studies in Agency and Responsibility vol. 3. Oxford: Oxford University Press.

Smart, J.J.C. 1961. “Freewill, Praise, and Blame.” Mind 70, 291-306.

———. 1973. “An Outline of a System of Utilitarian Ethics” in Utilitarianism: For and Against. Cambridge: Cambridge University Press.

Strawson, P.F. 1962. “Freedom and Resentment” in Free Will. Ed. D. Pereboom. Indianapolis: Hackett, 2009. 148-171.

Tadros, V. 2011. The Ends of Harm. Oxford: Oxford University Press.

Van Straaten, ed. 1980. Philosophical Subjects: Essays Presented to P.F. Strawson. Oxford: Clarendon Press.

Vargas, M.R. 2013. Building Better Beings. Oxford: Oxford University Press.

Wallace, R.J. 1994. Responsibility and the Moral Sentiments. Cambridge, MA: Harvard University Press.

---

[1] Vargas and others refer to such theories as 'moral influence theories' (2013, 166). I label them 'instrumentalist' to distinguish them from theories that explain responsibility in terms of influenceability, and which may or may not be instrumentalist.

[2] But see Hieronymi (2019) for an argument that blame is non-voluntary.

[3] See Eggleston for other definitions (2014, 134). See Miller (2014) for contrasts with rule and motive consequentialism.

[4] McCloskey (1965) gave an early formulation of this worry, discussed by Smart (1973, §10).

[5] Some argue that desert is a type of fittingness that bears directly on the all things considered appropriateness of blame, rather than indirectly on the moral permissibility of blame. See Shoemaker for a discussion (2020, 221-222).

[6] Arneson labeled Smart's theory 'influenceability.' I use '*' to distinguish my formulation.

[7] This is different from the charge that the moral psychology presumed by act consequentialism is either incoherent or unrealistically compartmentalized. Dale Miller (2014) gives a Strawsonian defense of Hare's utilitarianism against such a charge.

[8] H.L.A. Hart seems to have introduced the phrase 'economy of threats' to describe Benthamite utilitarian justifications of punishment, which he found unsophisticated (2008, 42).

[9] Jacob Barrett (2020) makes the same point. He agrees that classic moral influence theories were unsophisticated, but argues that they can incorporate Strawsonian insights. However, Barrett's aim is to defend instrumentalism and he doesn't offer a responsibility theory.

[10] Indeed, few such theories offered *any* account of how and how not to influence.

[11] This worry is different from another dependence worry that Vargas discusses. Here the concern is about the dependence of one's *responsibility theory* on one's moral theory. However, Vargas also thinks our *justification of the practice of blame* should be independent of our preferred moral theory. That is, for Vargas, his view that our blaming practices are "justified by their role in cultivating a form of agency sensitive to moral considerations" can be paired with different moral theories, which will give different answers to the question of what counts as a moral consideration (2013, 184-185)—e.g., outcomes, promises, roles/relationships, etc.

[12] Not only facts about today matter. McGeer seems to accept that Simon must be causally responsible for the offense. However, she denies that being reasons-responsive is a property that one can have *at a time*, so it cannot be a necessary condition on his responsibility that Simon was reasons-responsive *yesterday*. Rather an agent is responsible in virtue of being "sensitizable to considerations one failed to be sensitive to before" (2015, 2647). As I read McGeer, this a condition Simon must meet today (at the time of blame) not yesterday (at the time of the offense). Admittedly, she is not explicit about this last point and McGeer and Pettit seem to think that Simon must have the relevant capacity at the time of the offense (2015, 185-186).

[13] Thanks to Sofia Jeppsson for these examples.

[14] Thanks to Mark Alfano, Olle Blomberg, Daphne Brandenburg, Anneli Jefferson, Yuliya Kanygina, Benjamin Matheson, Theron Pummer, and Philip Robichaud for helpful comments on earlier drafts of the paper. And thanks to audiences at the VU Amsterdam, University of Oslo, University of Gothenburg, and the University of Zagreb for feedback and guidance early on. This paper was completed with funding from the Swedish Research Council (grant #2018-01156).