Forthcoming in *Surrounding Self-Control*, ed. by Al Mele (Oxford UP)

Negligence and Social Self-Governance

Manuel Vargas
University of California, San Diego

One way to understand self-control is in terms of something like inhibitory control, or more broadly, as a kind of self-regulation in the face of impulses that can distract or tempt the agent. Self-regulation of this sort is an important form of self-control. There are, however, broader notions of self-control that are worth the name. Among the most important of broader notions of self-control are one that figure in attributions of responsibility, or cases where the self is called to account for putatively culpable wrongdoing. The kind of self-control implicated in responsibility practices can sometimes involve inhibitory responses—for example, avoiding distraction and managing wayward impulses—but it typically requires much more than that. Paradigmatically, it implicates capacities to anticipate, to recognize reasons, or to call to mind considerations that may not be obvious. Even so, a number of philosophers have noted that there is something deeply puzzling about self-control of this more robust sort: it seems to require control over things for which we do not obviously have control at all.

Cases of recognitional failures abound. Consider cases where one screws up by forgetting to go to a meeting, or when one fails to notice a friend's new hair color, or when one unintentionally overlooks a clear defect in a proposal. These failures to notice, remember, or perceive are sometimes the basis of culpability. For accounts that account for responsibility in terms of self-control, or in the active exercise of agency, it is not obvious how this could be the case. These failures to notice, remember, and perceive aren't obviously instances of active, controlling agency. So, something seems amiss.

Section one considers these cases—*non-volitional culpability* cases—and explores how and why these cases are troubling for many accounts of responsibility. Section two presents a picture of the normative basis of moral responsibility—the *agency cultivation model*—and shows how this picture can resolve the puzzle in an appealing way. The kind of control that matters for responsibility is a product of the interaction of psychological and social pressures. It is a kind of socially-scaffolded self-control that appeals to a special notion of an agent's capacities. Section three argues that this account also provides a kind of reconciliation, or perhaps dissolution, of a particular debate in the literature on responsibility between reasons-responsiveness and attributionist approaches to moral responsibility.

1.

On one way of thinking about moral responsibility, direct (i.e., non-derivative) responsibility requires an important form of self-control, and in particular, control over one's actions and willings. As a first pass, we might say that an agent has direct control over some act *A* when he or she has the ability to voluntarily perform (or refrain from performing) *A,* on the basis of considerations that he or she recognizes as pertinent to *A*.

The picture has several appealing features. First, it has an elegant symmetry in its handling of actions and willings. My *action* is under my control when what I do stands in the right relationship to the considerations I recognize. My *willing*, or intention formation, is under my control to the extent that it suitably reflects the content of the considerations I recognize. If we think of control as the hallmark of a philosophically important notion of freedom, we might even say that this account gives us a unified picture for both free action and free will.

A second virtue of the approach is that it easily accommodates a thought once expressed by Dennett, that the appeal of construing freedom in terms of rationality "is expressed in the literature so often that there is probably something deeply right about it" (1984, 21). On this way of supplementing the basic idea of control, responsible agency is to be understood primarily in terms of a capacity to recognize and respond to reasons. This species of control-focused accounts can be called *Reasons* accounts of moral responsibility. Although there are important differences among Reasons accounts, they share a commitment to the idea that the species of freedom or ability most relevant to moral responsibility is properly connected to our rational powers (Wolf 1990; Wallace 1994; Fischer and Ravizza 1998; Nelkin 2011; McKenna 2013; Vargas 2013a). Control-based accounts, especially in the form of Reasons accounts, provide an intuitively appealing fit between, on the one hand, pre-philosophical convictions about control and, on the other hand, philosophical commitments to the centrality of rational capacities in responsibility.

Control-based accounts, fleshed out as Reasons accounts, allow for at least two distinct ways of avoiding culpability. First, an agent can fail to be responsible because of some lack of information. If through no fault of yours, you did not know that I am standing behind you, you are not culpable for turning into me as you turn around. Ignorance excuses, when it does, because it shows that the agent lacked access to the considerations necessary for the relevant sort of control (the recognitional component of control). Second, so-called "volitional" impairments, or failures of responsiveness to the relevant considerations, can also excuse. Consider the thought that, at least sometimes, some impulses can be extraordinarily difficult for particular agents to resist, even when those agents are alienated from the impulses, and even when acting on them comes at enormous risk to themselves and their wellbeing. Control-based accounts can hold that in these cases the agent lacks sufficient control for full responsibility. Of course, one might go on to insist that the agent is derivatively fully responsible if, for example, there was some prior culpable failure to expunge, restrain, or suppress that urge. Inhibitory control can and does sometimes matter a good deal for responsible self-control. Even so, the hallmark of control-based accounts is the thought that non-derivative responsibility is to be understood in terms of something like rational self-control.

Control-based accounts face a serious challenge from a class of apparent counterexamples, of which negligence is perhaps the most important.[1] Where recklessness involves something like a disregard for considerations that are apparent, negligence seems to involve a failure of awareness or regard that we think *ought* to have been present. If one forgets the fifth decimal place of pi, we ordinarily do not regard this failure as negligence. If someone fails to appear at conference because it slipped his mind to make travel plans in due course, we tend to regard that person as having failed in a notable way. In the language of legal theory, the offender has failed to meet a requirement of due care.

It may seem unclear how to accommodate our attitudes about negligence in a control-based account. One's neglect of the morally preferable option is, in the ordinary case, non-volitional. We do not intend to fail to consider some salient consideration, we do not plan to ignore demands for due care, and we do not usually seek to forget about shared commitments. We just forget to check whether the child was in the car, or that it is our anniversary, or that we

[1] A different line of attack, which I will not address here, has focused on whether the control condition invites a vicious regress. For a classic statement of the concern, see Strawson (1994). Others have noted that the control condition on responsibility is the chief source of worries about moral luck (Enoch 2010).

have plans to meet a colleague for coffee.

The moral significance of cases of non-volitional culpability seems to depend on the fact that there was a clearly morally preferable alternative to which the agent was not alive at the time of the offending act (or omission). It is the fact of the availability of that alternative, it seems, that gives rise to the charge of culpability. So, it looks like the control theorist has to show how agents could be alive to these alternatives. Yet failures of recall and recognition seem to failures to be alive to the relevant alternatives. It is therefore unclear how the putatively culpable agent could be in control in a way that grounds responsibility.

One intuitively appealing response is to insist that the agent could have remembered. There is something right about this response, as we will see. Again, though, it is not obvious why. Although I can sometimes succeed in recalling some bit of information on demand, I cannot make myself remember all the relevant facts at all the relevant times.

There are other cases of non-volitional culpability, cases that suggest that the relevant category is broader than culpable unwitting omissions. [2] Beyond failing to exercise due care, and memory lapses, we are prepared to hold people responsible for recognitional failures and failures of *reaction.* More controversially, we can even find people culpable for the underlying disposition or character trait implicated in such failure.

If Juana performs a great favor at considerable personal cost, yet Annie responds with a sense of (unearned) entitlement, we tend to think Annie is ungrateful. Interestingly, we need not think her ingratitude is a matter of direct volitional control. The mere fact that someone is

[2] George Sher (2006) provides a number of vivid illustrations of the moral stakes in failing to meet the care-taking demands of particular social roles, and his account is particularly effective at illuminating some of the philosophical stakes.

ungrateful, even as an involuntary reaction to great generosity, seems sufficient for us to find someone culpable. The observation generalizes: we find fault in involuntary indifference, apathy, or unintentional displays of attitudes of personal entitlement even when these things seem not to be matters of direct volitional control (Smith 2005; Raz 2011).[3]

Cases of non-volitional culpability are neither isolated nor rare. They are thus serious problems for control-based accounts. Either of two lines of response may seem promising for the Reasons-style control theorists: a *tracing* response and a *capacitarian* response. The tracing response insist that any cases of non-volitional culpability is derivative, or that it can be traced back to some prior culpable failure of control on the part of the agent. The capacitarian response holds that cases of apparent non-volitional culpability can be accounted for in terms of the agent possessing (but failing to exercise) a rational capacity, the possession of which grounds culpability.

On the face of it, both strategies face significant difficulties. An immediate difficulty for the tracing strategy is that our culpability judgments are frequently reactions to a *proximal* defect of the agent, not some distal, antecedent defect (McKenna 2012, 191; cf. Adams 1985, 14). We can, of course, extend the scope of agency by setting plans and policies for ourselves (Bratman 2000). The efficacy of our planning agency depends, in part, on how good we are at anticipating the need to do such things. However, the ability to anticipate such needs in advance of instructive prior experience seems to be vulnerable to just the same problem the idea of derivative responsibility was intended to solve. If I do not know that I am going to need to have elaborate

[3] I say "non-volitional" rather than "involuntary" or "non-voluntary" only because it is unclear whether the scope of the involuntary or non-voluntary properly extends to reactions, omissions, and inadvertence that one might approve of, or counterfactually endorse or with which one would identify. Natural language does not seem to decide whether it makes sense to call something involuntary if one did not consciously intend to perform the action, but one nevertheless identifies with the motives and values that lead to the action. In contrast "non-volitional" seems a clearer fit with instances in which one did not intend the act, however much that act or omission might cohere with the agent or the agent's idealized self.

reminders to remember my first anniversary, it will be much harder for me to be aware that I need to institute some or another mechanism that ensures that I will not forget. Appeals to derivative responsibility are limited by our "epistemic radar," as Michael McKenna has put it (2012, 191). Tracing a given instance of derivative responsibility back to some occasion of original responsibility succeeds, when it does, only when the case of original responsibility is one in which the agent can anticipate the consequences (or, perhaps, the kinds of consequences) that follow. However, downstream dispositions and non-voluntary reactions are not always among the things apparent to the agent at the time of original responsibility (Vargas 2005).

The capacitarian response seems more promising. Phenomenology, though, does not appear to be on the side of the rational capacitarian. Although we may want to say that I could have noticed my friend's freshly coiffed "do," it does not always feel like I could have. The capacitarian may need to dismiss phenomenology to save the normative metaphysics by holding that what matters is whether putatively negligent agents in fact have an unexercised capacity to recognize and respond to the relevant considerations. When there is an unexercised capacity to recognize the relevant consideration, it is true that non-volitionally culpable agents could have recalled (or could have been aware of) the relevant consideration, without a corresponding feeling or sense of that capacity. Such an account would be committed to a view according to which agents can have responsibility-supporting control in cases where they are unaware of the relevant consideration, and are unaware of their capacity to detect the relevant consideration. This is the approach that will be pursued in what follows.

It seems obvious enough that people have unexercised capacities. So, one might suppose, the capacities to recognize and respond to considerations might be understood in just the same way, whatever that proves to be. Yet, it will not do for the control theorist to deflect every potential counterexample by baldly asserting the existence of an otherwise unargued for capacity. Moreover, saying what such a capacity consists in seems to require a trip through the minefield of debates about what it means to say someone can do something other than what they do.

These challenges to the capacitarian construal of the control-based approach can be met, but they require grappling directly with the question of which powers matter for culpability, and why those instead of some others.

2.

Elsewhere, I have developed what I call the *agency cultivation model* of moral responsibility (Vargas 2013). For present purposes, a quartet of important ideas from that account will figure in the account of the sense in which non-volitionally culpable agents had an unexercised capacity to respond to suitable reasons. These four ideas include a presumption of our (imperfect) rationality, an account of the normative character of responsibility practices, the idea of social self-governance, and distinctive way of understanding the kind of capacity at the heart of a culpability assessments.

At least sometimes, we respond to reasons or considerations.[4] Whatever it is that constitutes our being able to recognize and suitably respond to reasons, we can be better and worse at it. (Let's be agnostic about the fundamental ontology of reasons, and stick with familiar social,

[4] While there is disagreement concerning how best to characterize the power of being able to recognize and respond to reasons, whether we have it in greater or lesser frequency, and whether such a power is sufficient to support our responsibility characteristic practices and/or ascriptions of free will, there is little disagreement with the claim that we are at least sometimes rational. To be sure, there is serious disagreement concerning how to understand what reasons are, and the proper characterization of the relationship between reasons, reasoning, and our affective or emotional states and dispositions. Nothing here turns on those disputes.

practical, and theoretical senses of "different kinds of reasons.") Not everyone is maximally good at recognizing the force of, say, mathematical, sartorial, sporting, and moral considerations. Our capacities to recognize and respond to one set of considerations do not seem to guarantee a comparable ability across another class of reasons.

We tolerate wide variances in the ability to recognize and respond to different kinds of reasons. You might be adept at recognizing the reasons to adopt a particular tempo in a given concerto, while I might be entirely oblivious to such reasons. I might be comparatively efficient at recognizing a reason to pass the soccer ball backwards, and you might make out such reasons only in rare circumstances.

Nevertheless, there are some forms of responsiveness that play a special role in shared, cooperative life. In particular, our responsiveness to moral considerations plays an especially significant role in our interpersonal practices. It is a partial basis for moralized praise and blame, in that the ability to recognize and respond to moral considerations constitutes our being responsible agents, or agents that are properly subject to norms of moralized praise and blame. What these practices, judgments, and attitudes express is, among other things, a demand that agents conduct themselves in some ways and not others, typically in ways that involve a certain minimal threshold of recognizing the reasons-giving force of other people's interests and values. So, if responsibility practices are to be what they present themselves as being, it is not enough that we find ourselves with them. They must be *justified*, at least if we are to retain them in anything like their current form.[5]

[5] Our psychological dispositions, our interests, and the things we find ourselves regarding as morally salient obviously shape our practices. Those dispositions provide a kind of constraint, internal to the practices, on what patterns of behavior we can demand and reasonably expect adherence.

This thought brings us to the second element of the account, a story about what would be minimally sufficient for justifying practices like ours. The core idea is this: when we hold responsible moral considerations-responsive agents (minimally, when we evaluate them in culpability-entailing ways) we participate in a system of practices, attitudes, and judgments that support and improve our responsiveness to moral considerations. These practices target agents that have a certain threshold of ability to recognize and respond to moral considerations. Over time, and given psychologies roughly like ours, praise and blame and the related apparatus of responsibility practices performs an important function for us. That is, they sustain and further develop those moral considerations-responsive capacities that seem to naturally occur wherever groups of humans are to be found.

If the foregoing is correct, responsibility practices earn their keep in our social lives by fostering and sustaining a particularly valuable form of agency. Such accounts sometimes raise the worry that agents cannot be properly said to merit or deserve these moralized reactions. However, a teleological account of responsibility's justification does not entail that individual judgments of responsibility are not backward-looking, desert-entailing judgments (Rawls 1955; Hart 1959). The teleological element—the imperative to build better beings—is a feature of *the system* of first order norms. Desert judgments are judgments at the level of the first order norms and not in conflict with the second-order teleological character of the account (Doris 2015; Vargas 2015, 2019). As long as these (desert-entailing, maybe retributive) norms and practices have the right systemic effects for creatures like us, we have some justification for a system of responsibility practices.[6]

[6] On this account, one can accept that retributive attitudes have a role in our moral life, without also endorsing

The third important feature of the approach is the idea of social self-regulation, or the idea that our ability to self-govern is partly dependent on social scaffolding of our moral lives from without. Consider the characteristic way in which our practices interact with our psychologies. Acceptance of blame, in creatures like us, is typically marked by the experience of guilt. That guilt casts a pall over our estimation of our own actions. Guilt provides a motivational impetus, in agents concerned with either moral demands or social standing, to repair or restore those prior relations. If I go unblamed, it is harder for me to experience the guilt that motivates moral self-improvement, and to undertake moral repair with those I have wronged.[7] So, internalized blame norms and practices provide a widely-distributed source of social guidance for individual agency. Our self-governance, our control over intentions and actions, relies on a social feedback. More specifically, our moral considerations-responsive agency is sustained and enhanced by practices of judging and holding responsible.

The social interest in policing the moral failures of others interacts with a personal interest in our own agency. We have an abiding interest in distinguishing between those domains or circumstances where our agency is reliable and competent in its functioning, and those where it is not. In a word, the difference is control.[8] There are the easy cases, where it is manifestly evident that agency ceases to be efficacious in a reliable and competent way. Were I to fall from an airplane without a parachute, we might suppose that I will have little control over where I am going and the terror over my imminent demise would swamp my mental life, further shrinking

retributive criminal punishment as it is pursued in, for example, the United States.

[7] This is a view that has been developed in several places, including Bennett (2002) and McKenna (2012).

[8] Here, I draw from recent work of Joseph Raz (2011) on responsibility, although he characterizes the notion in terms of "domains of secure competence" rather than domains of control.

my agential powers in that context. For most of us, however, such radical narrowing of our powers is a relatively uncommon thing. Our lives tend to be characterized by a mix of slowly expanding and retreating domains of reliable competence, that is, of greater and lesser control.

Here, though, we can see where the social aspect of self-governance connects to an individual's own interest in self-governance, in general, and with respect to morality in particular. As individual agents, we have a robust interest in which actions in which context fall into our domain of reliable competence. However, some actions have a significance that is socially structured. Whether I am reliably competent at being a good scholar, a good teammate, an upstanding member of a religious community, and so on, is partly a matter of what the local, social constitution of those roles comes to. I have a reason to attend to the social dimension of those roles, and to structure my behavior accordingly, inasmuch as it matters to me to be seen as competent in those ways.

Failures of competence in this or that role may not always matter much to an individual agent. However, recognition by others of an agent's widespread incompetence across social domains comes at substantial cost to the agent. This is especially so when the agent proves to be incompetent at navigating a class of concerns that are ubiquitous and widely regarded as important, such as morality. Being seen as incompetent at navigating moral considerations is, minimally, to be marked as untrustworthy in a range of social relations. Indeed, if one does not cross a certain threshold of competence at moral considerations, one is typically not regarded as a peer, suitable for ordinary social relations characteristic of adults. Moreover, depending on the degree and scope of perceived moral incompetence, one can be regarded as an outright threat, as someone to be repelled or even eradicated from the social order.

It is all well and good that moral considerations-sensitive capacities can be significant for both social and individual reasons. It is another matter to say what those capacities consist in. So, I now turn to the final element: the specification of those capacities.

There are, we think, unexercised capacities. However, our naive or intuitive sense of whether we have a capacity or not does not always hold up well to empirical scrutiny about apparent exercises of those capacities (Doris and Murphy 2007; Nelkin 2005; Brink 2013; Vargas 2013b). So, we need an account of capacities that can support our responsibility practices, but in a fashion that does not seem patently implausible given what we find about actual human behavior.

One way to make sense of capacity talk is to invoke causal indeterminism as a central feature in the possession of a capacity. On this approach, people have some capacity—rational or otherwise—when, holding fixed the exact features of some circumstance and the laws of nature, there are two or more action possibilities available to the agent. (One might add to this some further requirement—many philosophers are skeptical that indeterminism is enough, even though some think that indeterminism is indispensable.) Suitably positioned in operations of agents, indeterminism secures the possibility of unexercised capacities, one might think.

There are various time-worn objections (conceptual, normative, and empirical) to this kind of view (Vargas 2013a, Ch. 2). The present approach takes its cue from a different idea. An unappreciated but striking feature of capacity talk is that in many contexts, the nature of a capacity is interest-sensitive. Whether I am capable of dancing a jig depends, in part, on your interest in asking, and what background assumptions we take to be operative in the question. Whether my being asleep matters for my ability to dance a jig depends on whether you are asking about my suitability for entering an Irish dance competition next week, or instead, whether your interest is in my entertaining you at that exact moment. In the former case, my sleep is no deterrent to my being able to dance a jig, and in the latter, it is. On this approach, the relevant metaphysical basis of my jigging capacity is partly picked out by the interests governing the question.

(We could distinguish between general and specific abilities, but this strikes me as artificial. The set of contexts over which capacities range almost always allows for greater and lesser gradations of granularity, in no small part because the *now* and the *then* that time-indexes our interests is itself often flexible.)

On this sort of approach, the immediate question to ask is this: *which* interests are the ones that properly govern the specification of the responsibility-relevant capacities, and why those and not others? The beginnings of an answer can be given by considering the idea of a *Sidgwickian capacity*. A Sidgwickian capacity is a capacity that is identified by an ideal observer, an observer with some specified interest or interests.

Let's suppose our ideal observer is in the actual world, fully informed, and ideally logical. We can specify the observer's interest by appeal to the aforementioned justification of the responsibility practices. Because those practices are justified by their effects in sustaining and enhancing moral considerations-sensitive agency, our ideal observer's interest is in selecting a notion of capacity that would be at least co-optimal for (1) ensuring that agents in the actual world recognize and suitably govern themselves in light of moral considerations, and (2) ensuring agents have wider rather than narrower ranges of context of action and deliberation in which agents so deliberate and act, so long as it does not conflict with (1). In short, the capacity that matters for culpability-entailing assessments of moral responsibility is one that in the actual world supports and extends our ability to recognize and respond to moral considerations.

In specifying the relevant Sidgwickian capacity, our observer is concerned with our current customary psychologies, the cultural and social circumstances of our agency, our interest in resisting counterfactuals we deem deliberatively irrelevant in the actual world, and the need for us to internalize norms of action and deliberation concerning moral considerations at a level of granularity that is useful in ordinary deliberative and practical circumstances.[9] Thus, an agent has the responsibility-relevant capacity if (a) he or she recognizes and appropriately responds to the relevant moral considerations, or, (b) if the agent recognizes and responds to moral considerations in a suitable proportion of relevantly similar worlds, as specified by the observer.

On this approach, whether someone has the relevant Sidgwickian capacity cannot be settled absent appeal to the normative aspirations of a system of responsibility, and various facts about existing practices. An agent's responsibility-relevant capacities are not something which neuroscience, psychology, or other putatively non-normative enterprises can settle on their own. Put differently, the facts about the relevant notion of capacity, and whether a given agent has the involved ability, is a "higher-order" or constructed fact, built on the basis of an agent's psychology, facts about context, and the normative structure of a justified system of moral responsibility.

(For ease of exposition, I'm suppressing various epicycles to the basic approach. These

[9] On the restriction to deliberative relevance in the actual world, think: finks and Frankfurt-style cases. In the actual world, Frankfurt cases are infrequent and not a possibility with deliberative significance in the ordinary course of things. In a world in which we had reason to think Frankfurt cases were common, it is conceivable that such cases could have a different significance for the relevant construal of capacities. For these and related esoteric delights of specifying the modal profile of the capacity see Vargas (2013a, 213-228).

include tools for addressing various worries about how to specify the worlds—for example, restricting them to worlds with laws like ours—or addressing how one counts proportions of plausibly infinite numbers of worlds (hint: maybe ideal observers are remarkably good at sampling arbitrarily huge but finite numbers of deliberatively relevant worlds?). For present purposes, we can put aside some of the metaphysical exotica and focus on the idea of a notion of capacity given by what is best for the functioning of a social practice. There are a variety of ways one might develop this general idea.)

We now have the tools to resolve the puzzle about involuntary culpability. There is no reason to suppose the Sidgwickian capacity will only give us exercised capacities, given the idea of social self-regulation of our moral agency. The social self-regulation picture is one that presupposes feedback from our responsibility characteristic attitudes and practices. One way we extend our capacities into new contexts is to, at some point, be vulnerable to blame because we had a capacity that went unexercised. Nonvoluntary culpability is, on this picture, a case of being able to recognize and suitably respond to some set of moral considerations centered on due concern for others—but failing to do so. Put differently, the agent had all the capacities required for responsibility-relevant control, but suffered from a failure to exercise them.

The notion of ability here will not map on to a notion of ability concerned with accessible worlds, as allowed by the actual past and the laws of nature. But that latter notion of ability turns out to not be required to support the social and normative concerns that structure responsibility. In contrast, the Sidgwickian capacity I have identified is plausibly required to play the requisite function in a system of moral responsibility practices, attitudes, and judgments. Both capacity and culpability are intimately bound up with the value of a form of self-governance, and our interest in our agency being a certain way. To be sure, there may be some causal etiology to a given failure to exercise the capacity. Unless that etiology works through familiar excuses, the presence of a causal explanation (that otherwise has no significance for the practice of responsibility) will be beside the point.

From the standpoint of individual agents, one's status as a morally competent agent is something to be fostered and monitored both for its own sake, but also because it is a precondition for full participation in most mature forms of human social activities. Agents with grossly unreliable competence at navigating moral demands are typically not recognized as full members of the moral community. It is thus unsurprising that we ordinarily seek to accurately track and maintain moral competence as best we can. In doing so, we relying on our own assessment, but crucially, on the assessment of others as well. In this way, an individual persons' concern for his or her domains of reliable control comes to be systematically interwoven with the collective, norm-structured demands of social life.

The agency cultivation model of responsibility provides an explanation of the conditions under which an agent who commits an act of non-volitional culpability could have been aware of a relevant consideration when she failed to actually be aware of it. This is not to deny that there is, in some sense, a way of ordinarily speaking according to which it was *not* in the agent's control that he or she could remember or become aware of the relevant consideration. This fact is compatible with there also being some other sense in which the agent could have remembered, could have reacted differently, and so on. A special version of this latter sense is what I have been endeavoring to specify.

In light of the foregoing, it is tempting to say that the present account is ultimately a vindication of the Reasons approach. What is central to Reasons approaches to responsibility was never the idea of capturing any and all notions of control. Rather, the aspiration was to identify a form of rational agency that could support the normative integrity of our practices, or practices very much like our current ones. The idea of agents having an interest in domains where their rational capacities are reliable, and the development of this idea in the distinctive, social self-governance way, yields a picture at some remove from the image of individualistic agents processing reasons in the absence of communities or other agents. Our evaluation of agents and their choices is instead a socially-embedded activity with the rules settled by various social and normative considerations.

3.

I conclude with some reflections about how a failure to appreciate the resources of Reasons accounts for handling non-volitional culpability cases have misled some recent work on the options available for theories of moral responsibility.

In the contemporary philosophical literature, non-volitional culpability cases are supposed to be the basis of a dispute between "volitionalists" and "attributionists." Volitionalists are those who identify the exercise of conscious, intentional control as the centerpiece of moral responsibility. In contrast, attributionists are those who reject such requirements, and instead ground moral responsibility in appraisal of some privileged attitudes of the considered agent. I am inclined to think that this putative dispute, the framing of it in these terms, has proceeded from a mistaken view about the theoretical options.

Over the past twenty years or so, "attributionist" accounts have come to be regarded as a serious competitor to conventional, control-focused accounts of responsibility. A bit of history may be in order. One important point of departure for attributionist accounts is an essay by Gary Watson (1996), wherein he distinguishes between two "faces" or aspects of responsibility. The first "face" is what he called *attributability*, the second is *accountability*. As the distinction is frequently cast, attributability concerns the "aretaic" face of responsibility, the face concerned with moral qualities or character of an agent, qualities that attribute a moral fault with the agent. The *accountability* face characteristically concerns culpability for actions, the sort of thing that gives rise to moral praise and blame.[10]

As the literature subsequently developed, it became customary to characterize a growing number of accounts as "attributionist." This is unfortunate for several reasons. First, it is unclear whether so-called attributionist accounts are unified in any deep way. Rejection of a control condition is merely a negative characterization, and provides little in the way of substantive commitments on how non-volitional culpability is to be understood. This is especially so if it is not clear that control-focused accounts share a unified conception of control. Second, among attributionist views there are deep disagreements about the basic issues and terrain.

Scanlon, for example, has insisted that moral responsibility concerns whether we can attribute an action to an agent, whereas blame involves a modification of one's relationship with

[10] To my ear, characterizations of attributability as moral responsibility, or as a form of moral responsibility disconnected from blameworthiness, stretch the standard meaning of 'moral responsibility'—which is not to deny that attributability picks out something important and distinctive. There is a wide spectrum of views about this matter, though. For example, McKenna is ambivalent about whether to treat accountability as a form of responsibility, sometimes accepting Levy's (2005) claim that responsibility in the attributability sense is just a way of noting either bad agency or "mere moral agency" (McKenna 2012, 185 n. 9). In contrast, Shoemaker (2011), Vincent (2011), and Fischer and Tognazzini (2011) are more permissive, treating attributability as a recognizable sense of 'responsibility'.

another on particular grounds (2008, 128). This is a view on which there is more to responsibility than blameworthiness and its requirements. In a similar vein, Smith cautions that "it is very important not to conflate claims about responsibility and claims about blameworthiness" (2005, 266). Raz has similarly claimed that "the preoccupation with praise and blame, natural in our blame society, misses the central role of responsibility" (Raz 2011, 265).

However, not everyone ordinarily regarded as an attributionist seems committed to separating blame from other forms of moral evaluation. Adams, another progenitor of the current attributionist efflorescence, insists that "it seems strange to say that I do not blame someone though I think poorly of him . . . thinking poorly of a person in this way *is* a form of unspoken blame" (21). Such a picture suggests an expansive notion of blame, applying to nearly any case of negative moral evaluation directed at agents, ranging from aretaic assessments to assessments of moral worth to judgments of culpability. If any negative aretaic evaluation either constitutes or entails blame, then the distinction between attributability and accountability threatens to collapse.[11] At the very least, there are distinct ways in which the "aretaic" or characterological face of responsibility has been understood. If Adams is right, though, the relationship of "attributability" to accountability, or culpability-imputing blaming practices is elusive.

It helps to be clear that there are plausibly a diverse range of evaluative attitudes implicated in accounts that are characterized as theories of responsibility. It is more promising to hold that there is a spectrum of evaluative attitudes, and that theorists are using the same label

[11] For Adams, punishment is distinct from mere reproach, in that punishing is only appropriate or applicable in cases where the agent acted voluntarily. So, for Adams' account, volitional requirements are only important for punishment.

for different parts of that spectrum. To anticipate: on one end of the spectrum, there are judgments of pure "grading," i.e., assessments of better and worse that describe features of an agent, without imputing culpability (Smart 1961). On the other end of the spectrum, there are culpability-imputing, desert-entailing judgments of blame and punishment. Much of the apparent disagreement between attributionists and their critics reflects distinct foci concerning the part of the spectrum that is at stake, and sometimes, disagreements about how best to label one's preferred part of the spectrum. Some of the confusions about labels is simply a byproduct of the fact that there is little convergence among theorists about which subset of that spectrum is proper to moral responsibility, as opposed to some other normative category.

It is again instructive to focus on Robert Adams' account. Adams' remarks suggest the expansive view that where there is negative evaluative assessment, there is responsibility. There is reason to resist this picture. One can form evaluative judgments ("Jim is bad at mathematics") without this entailing a commitment to culpability or deservingness of blame. Moreover, as Slote (1990) has suggested, characterological judgments—that one is generous, hateful, or courageous—can persist, even if we conclude that moral responsibility does not exist. One implication of this sort of view is that abandonment of ordinary forms of moralized praising and blaming for actions need not entail a complete loss of moral vocabulary. Indeed, several philosophers have argued that there is considerable diversity of moralized attitudes we can take towards our intra- and interpersonal lives, even were we to accept that culpability-bearing judgments that entail desert of praise and blame cannot be sustained (Pereboom 2001; Honderich 2004).

The view suggested by Adams' remarks, however, seems to suggest that aretaic evaluations entail criticism of those with characters or dispositions we describe in negative evaluative terms, and that blame follows in the wake of those negative evaluations. On this account, it does not matter whether the involuntary attitudes were intentionally acquired or whether their acquisition was under the control of the agent.[12] Adams view, then, seems to be one on which the category of attributability absorbs what was supposed to be distinctive of accountability, i.e., blame.

The view I am ascribing to Adams appears to run together two things that are often compresent but distinct. On the one hand, there is a disposition we can label in normative language—say, calling something *vicious*—and second, there is an assessment whether that feature reflects on the agent in some creditable way (whether for good or ill). Adams seems to think that any aretaic or agent-targeted version of the former automatically entails the second.

The view suggested by Smart's remarks, however, is more plausible. Dogs and human might both be vicious, but ordinarily we presume that only one can be called to account for viciousness. Dogs might not be the same kinds of agents that adult humans characteristically are, but it suggests that the proper way to think about at least some aretaic appraisals is that they mark out normatively significant dispositions. It is a further matter whether and how the agents with those dispositions can be called to account for them, or whether in having those dispositions they deserve blame.[13]

[12] His model is responsibility for beliefs. As he sees it, the beliefs of a member of the Nazi Youth organization are, regardless of origin or their conditions of acquisition, fitting targets of blame: "No matter how we he came by them, his evil beliefs are a part of who he is, morally, and make him a fitting object of reproach. He may also be a victim of his education; and if he is, that gives him a particular claim to be regarded and treated with mercy—but not an exemption from blame" (19).

[13] Indeed, normatively salient dispositions are not even limited to agents. We might offer evaluative judgments of states of affairs that entail no deservingness of praise or blame but merely mark out differences in value. Whether or not people deserve disapprobation for what they say and do, we can still regard it as an inferior world if people are cruel, access to the goods of political life are unequally distributed, and good deeds are met with bad. Grading is not restricted to normatively labeling of characters or dispositions, then. It may also be applied to other ways we find

So, should we understand attributionist accounts as either mistaken about the necessary import of aretaic and similar normative assessments, or alternately, as collapsing the distinction between attributability and accountability, i.e., deservingness of moralized praise and blame?

The answer seems to be neither. Some self-described attributionists have argued that they offer an account of moral responsibility that is neither merely aretaic, nor centrally concerned with deserving blame. Angela Smith's (2005, 2012) attributionist account is perhaps the best-known version of such an account. She explicitly disavows a concern with "merely" aretaic assessment. For Smith, the core issue of responsibility seems to be whether the considered feature of character, the reaction, or the mental state of the agent is connected to that agent's judgments in such a way that the agent can cite the reason for it. For this reason, this version of attributionism is taken to focus on a property we can, following Shoemaker, characterize as *answerability*.[14]

There are some underappreciated difficulties for a Smith-style account. Her presumption is that non-volitional actions reveal evaluative attitudes. However, there are a variety of empirical considerations that cut against the presumption. First, there is the large body of experimental psychological research that suggests a range of actions, both voluntary and involuntary, reflect situational triggers, cultural scripts, or the reproduction of statistical estimates in a way that agents can and do sincerely disavow (Doris 2002; Nelkin 2005; Brink 2013; Vargas 2013b).

the world, whether social organizations or patterns of actions and events.

[14] David Shoemaker (2011) distinguishes answerability from attributability and accountability. Smith seems to largely accept that the notion she previously characterized in terms of attributability is the same notion that Shoemaker identifies as answerability. Because of this, it is unclear whether McKenna's contention that Smith collapses accountability and attributability can now be sustained. Presumably, Smith would insist that answerability and accountability *can* be distinguished. However, McKenna notes that on Smith's view, blame just is moral criticizability (Smith 2008, 377), which makes her account much closer to Adams', at least with respect to blame.

Depending on the frequency that such influence have—an empirical matter—it may be that ordinary instances of action are poor (or even not at all) evidence of the agent's evaluative commitments. One could, of course, modify the view in various ways, argue for non-conscious values, and the like. Even were we to have such an account, accepting them over alternatives would remain a tendentious matter.

Second, there are other phenomena, such as mood disorders, which suggest that the relationship between avowals, an agent's privileged bits of psychology, and the question of what speaks for the agent will be extraordinarily complex and not so clear cut as to allow us to read off an agent's evaluative commitments from actions. There is some pressure in ordinary thought to hold that actions derived from mood disorders—say, depression, hypomania, or seasonal affective disorder—obscure rather than reflect the agent's evaluative judgments. The idea is that these disorders do not alter an agent's "true" or "real" evaluative commitments, but rather, they mask them. These are, of course, difficult issues in how we understand the idea of selves and the operations of various disorders. One can, for example, insist that the hypomanic person's characteristic over optimism and inconsideration of others is no obscuring of the privileged commitments or evaluative attitudes, but rather a change in those commitments and attitudes. This is at odds with the idea that such disorders make people "not themselves." So, on pain of giving up the transparency of an agent's evaluative commitments, it seems that Smith must maintain that the depressed person, the hypomanic person, and so on, are indeed being themselves when they are depressed, manic, or otherwise disordered.

Third, there is the fact of performance errors, or slips (Amaya 2013). Grammatical errors and pronunciation glitches are common in ordinary speech, even among fully competent speakers of a language. More generally, attentional and volitional glitches appear to be relatively common in agents. As Amaya and Doris put it, "the mistakes people make are not a reflection of deep seated attitudes in them but are rather due to small lapses of concentration and memory. Interestingly, these *slips* do not happen randomly but come in systematic patterns, which suggests that they are not isolated glitches but rather mistakes in the life of altogether normal agents" (Amaya and Doris 2015, 264). The crucial point here is that these errors do not reflect a lack of moral concern on the part of the offending agent. Indifference, obliviousness, and callousness are not the culprits in actional failures any more than they are in grammatical failures. Instead, it is our human, all-too-human nature. To the extent to which such failures are as common as Amaya and Doris suggest, it seems hasty to conclude that we can readily read an agent's non-volitional failures as evidence of their evaluative commitments.

A fourth and different reason to resist Smith's picture is that answerability sets a high bar. We do not always know why we act. We confabulate with some regularity, and in general, our psychology and values are not always evident to us (Wegner 2002; Arpaly 2003; Knobe and Roedder 2009). To insist that the core of responsibility is answerability seems to make responsibility an overly-intellectualized thing, the domain of psychologists and philosophers rather than ordinary people navigating social space.

To her credit, Smith's views have complexified in ways that have closed the gap between her view and a Reasons-style view. In this context, it may be worth reflecting on the fact that a good deal of the initial impetus for developing the attributionist approach as a distinctive approach was explicitly rooted in the purported inability of control-focused accounts to explain culpability for non-volitional action. This effort turned on a failure of the literature to have made clear what resources rational capacitarians had for accounting for these cases. If this is right, the development of attributionist approaches was a response to an illusory problem.

Skepticism about capacitarian strategies has helpfully focused our attention on the conditions required for agents to explain their acts, and to defend their motivations and commitments. Moreover, reflecting on what these accounts are committed to has given us a better appreciation of the varied terrain surrounding responsibility. As noted above, for example, Shoemaker (2011) has argued that we need to distinguish when a morally salient quality applies to an agent (attributability), when an agent can properly be called to answer for his or her attitudes (answerability), and accountability, or when an agent is culpable for attitudes or actions.

These taxonomical insights are important. For example, given the range of things that have been labelled responsibility—judgments of grading, aretaic assessments, the ability to give reasons for one's actions, acts and omissions that support praise and blame, and so on—one suspects that the term "moral responsibility" has become a vehicle for more confusion than illumination. At the same time, those who have endeavored to offer accounts of responsibility that are explicitly not reliant on notions of control, or who have otherwise offered accounts that tend to be thought of as attributionist, do not seem to share a sense a unified picture of what evaluative attitudes are central and/or what it is that distinguishes attributability from other notions widely taken to be central to moral responsibility.[15]

How one chooses to label things is, of course, a matter of preference. I prefer to restrict talk of moral responsibility to matters centrally connected to moralized blameworthiness and praiseworthiness (i.e., accountability). On this way of arranging terminological preferences, when

[15] It may be that particular theories will show that one or more of these categories ultimately collapse into another, but that would be a discovery and not something we should build into our initial construal of the landscape.

we use ‘moral responsibility’ as a label for attributability and answerability—but also authenticity, autonomy, and other idealized notions of self-governance—we needlessly invite confusion. This is not to say that accounts employing permissive usages of ‘moral responsibility’ are themselves introducing confusions. It is only to say that we should be wary of talk of responsibility that does not clarify the judgments, attitudes, and practices that are central to that account.

What is central to my account here has been the idea that failures to recognize and respond to relevant moral considerations can constitute a fault of the agent. The explanation for why is tied, in part, to failures to exercise our rational capacities in a sense that is tied to both our own individual interests as agents but also tied to the conditions of shared, cooperative life with other moralizing creatures. When we judge that someone has failed to self-govern in ways sensitive to the moral concerns that figure in our social world, we typically go on to blame then.

The old voluntarist/attributionist distinction does not carve the joints of the philosophical issue. Reasons-responsiveness views, especially of the *de re* responsiveness variety (e.g., Arpaly and Nelkin, among others), do not map on to any of the characterizations they give of voluntarist views, for they need not be focused on choices, especially of an exclusively conscious variety. What matters is whether the agent (or on some views, a particular sub-agential mechanism) is responsive in the right way to the relevant considerations. It need not matter whether that responsiveness goes through an explicit, conscious choice. Thus, at least some reasons-responsiveness views seem to be able to say a lot of what attributionists want to say. Indeed, Smith acknowledges that “If one thinks that ‘choices’ can, like judgments, be inexplicit, unconscious, and attributed to a person simply in virtue of her responses, then my disagreement with the volitional view would turn out to be much less significant" (2005, 256).

The relevant ground of culpability is not choice, but rational capacity. Once we allow that control can be had in light of a rational capacity, the operations of which the agent need not always be conscious of, and once we see how we can make sense of the idea of an unexercised capacity, and once we acknowledge that both individuals their communities have an interest in being prepared to blame agents with such capacities, the puzzle about non-volitional culpability evaporates. We can explain this as the triumph of the core idea of the Reasons view, or as a vindication of a suitably sophisticated volitionalism, or as a concession to attributionist concerns about non-volitional acts and omissions, or even as the collapse of a meaningful distinction between attributionists and volitionalists. Which we choose is mostly a matter of labels.[16]

[16] Prior versions of this paper were presented at U.C. Davis, the College of William & Mary, Sonoma State University, the University of Arizona, U.C. Irvine, and the 7th Annual Meeting of the AAMP. My thanks to those audience members for helpful feedback on the paper, especially Michael Bratman, Maite Ezcurdia, Michael McKenna, Al Mele, Agustín Rayo, and Stephen Wall. I'm also very grateful for extensive comments on drafts of this paper by Fernando Rudy, Santiago Amaya, and Samuel Murray. This chapter was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this chapter are my own and do not necessarily reflect the views of the John Templeton Foundation.

References

Adams, Robert Merrihew. 1985. "Involuntary Sins." *Philosophical Review* 94 3–31.

Amaya, Santiago. 2013. "Slips." *Nous* 47 (3): 559–76.

Amaya, Santiago, and John Doris. 2015. "No Excuses: Performance Mistakes in Morality." In *Handbook of Neuroethics*, edited by Jens Clausen, and Neil Levy, 253–72. New York: Springer.

Arpaly, Nomy. 2003. *Unprincipled Virtue*. New York: Oxford.

Bennett, Christopher. 2002. "The Varieties of Retributive Experience." *The Philosophical Quarterly* 52 (207): 145–63.

Bratman, Michael E. 2000. "Reflection, Planning, and Temporally Extended Agency." *The Philosophical Review* 109 (1): 35–61.

Brink, David O. 2013. "Situationism, Responsibility, and Fair Opportunity." *Social Philosophy and Policy* 30 121–49.

Dennett, Daniel. 1984. *Elbow Room*. Cambridge: MIT.

Doris, John. 2002. *Lack of Character*. New York: Cambridge University Press.

Doris, John. 2015. "Doing Without (Arguing About) Desert." *Philosophical Studies* 172 (10): 2625–34.

Doris, John, and Dominic Murphy. 2007. "From My Lai to Abu Ghraib" *Midwest Studies in Philosophy* 31 25–55.

Enoch, David. 2010. "Moral Luck and the Law." *Philosophy Compass* 5 (1): 42–54.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Fischer, John Martin, and Neal Tognazzini. 2011. "The Physiognomy of Responsibility." *Philosophy and Phenomenological Research* 82 (2): 381–417.

Hart, H. L. A. 1959. "Prolegomena to the Principles of Punishment." *Proceedings of the Aristotelian Society New Series* 60 1–26.

Honderich, Ted. 2004. "After Compatibilism and Incompatibilism." In *Freedom and Determinism*, edited by Joseph Keim Campbell, Michael O'Rourke, and David Shier, Cambridge MA: MIT Press.

Knobe, Joshua, and Erica Roedder. 2009. "The Ordinary Concept of Valuing." *Philosophical Issues* 19 131–47.

McKenna, Michael. 2012. *Conversation and Responsibility*. New York: Oxford University Press.

McKenna, Michael. 2013. "Reasons-Responsiveness, Agents, and Mechanisms." In *Oxford Studies in Agency and Responsibility, Vol. 1*, edited by David Shoemaker, 151–84. New York: Oxford University Press.

Nelkin, Dana. 2005. "Freedom, Responsibility, and the Challenge of Situationism." *Midwest Studies in*

Philosophy 29 (1): 181–206.

Nelkin, Dana Kay. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Rawls, John. 1955. "Two Concepts of Rules." *Philosophical Review* 64 3–32.

Raz, Joseph. 2011. *From Normativity to Responsibility*. Oxford: Oxford University Press.

Scanlon, Thomas. 2008. *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge, MA: Belknap Press of Harvard University Press.

Sher, George. 2006. "Out of Control." *Ethics* 116 (2): 285–301.

Shoemaker, David W. 2011. "Attributability, Answerability, and Accountability" *Ethics* 121 602–32.

Slote, Michael. 1990. "Ethics Without Free Will." *Social Theory and Practice* 16 (3): 369–83.

Smart, J.J.C. 1961. "Free Will, Praise, and Blame." *Mind* 70 291–306.

Smith, Angela. 2005. "Responsibility for Attitudes." *Ethics* 115 236–71.

Smith, Angela. 2008. "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138 367–92.

Smith, Angela. 2012. "Attributability, Answerability, and Accountability" *Ethics* 122 (3): 575–89.

Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75 5–24.

Vargas, Manuel. 2013a. *Building Better Beings*. New York: Oxford.

Vargas, Manuel. 2013b. "Situationism and Moral Responsibility." In *Decomposing the Will*, edited by Till Vierkant, et al, 325–49. New York: Oxford University Press.

Vargas, Manuel. 2015. "Desert, Responsibility, and Justification: Reply to Doris, Mcgeer, and Robinson." *Philosophical Studies* 172 (10): 2659–78.

Vargas, Manuel. 2019. "Responsibility, Methodology, and Desert." *Journal of Information Ethics* 28 (1):

Vincent, Nicole. 2011. "A Structured Taxonomy of Responsibility Concepts." In *Moral Responsibility: Beyond Free Will and Determinism*, edited by Nicole Vincent, et al., 15–35. Dordrecht, The Netherlands: Springer.

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

Watson, Gary. 1996. "Two Faces of Responsibility." *Philosophical Topics* 24 227–48.

Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wolf, Susan. 1990. *Freedom Within Reason*. New York: Oxford University Press.