Routledge
Taylor & Francis Group

Philosophical Explorations

An International Journal for the Philosophy of Mind and Action

ISSN: 1386-9795 (Print) 1741-5918 (Online) Journal homepage: http://www.tandfonline.com/loi/rpex20

Good intentions and the road to hell

Sarah K. Paul

To cite this article: Sarah K. Paul (2017) Good intentions and the road to hell, Philosophical Explorations, 20:sup2, 40-54, DOI: 10.1080/13869795.2017.1356354

To link to this article: http://dx.doi.org/10.1080/13869795.2017.1356354

Philosophical Explorations, 2017
Vol. 20, No. S2, S40–S54, https://doi.org/10.1080/13869795.2017.1356354
This paper is part of the online only special issue "The normative significance of intentions" guest edited by Ulrike Heuer and Matthew Noah Smith.

Good intentions and the road to hell

Sarah K. Paul*

Department of Philosophy, University of Wisconsin-Madison, 5185 Helen C. White Hall, 600 N. Park St, Madison, WI 53706, USA

(*Received 2 November 2016; final version received 24 March 2017*)

G. E. M. Anscombe famously remarked that an adequate philosophy of psychology was needed before we could do ethics. Fifty years have passed, and we should now ask what significance our best theories of the psychology of agency have for moral philosophy. My focus is on non-moral conceptions of autonomy and self-governance that emphasize the limits of deliberation – the way in which one's cares render certain options unthinkable, one's intentions and policies filter out what is inconsistent with them, and one's resolutions function to block further reflection. I argue that we can expect this deliberative "silencing" to lead to moral failures that occur because the morally correct option was filtered out of the agent's deliberation. I think it follows from these conceptions of self-governance that we should be considered culpable for unwitting acts and omissions, even if they express no ill will, moral indifference, or blameworthy evaluative judgments. The question is whether this consequence is acceptable. Either way, the potential tradeoff between self-governance and moral attentiveness is a source of doubt about recent attempts to ground the normativity of rationality in our concern for self-governance.

Keywords: agency; responsibility; omissions; deliberation

> Grahm felt increasingly frustrated. Voices in his mind told him to do something, but he had no ideas. Like most of his neighbors he had devoted his life to farming. He liked farming. All he wanted to do was farm. Farmers had a long, proud history of avoiding social, economic, and political issues. They enjoyed nature, work, and solitude, and they eschewed everything that might be considered grist for the nightly news. (David Rhodes, *Driftless*, 2008)

There is a significant strand of contemporary philosophy of action that has taken to heart Anscombe's (1958) proclamation in "Modern Moral Philosophy": that before we can profitably do moral philosophy, an adequate philosophy of psychology is needed. According to these conceptions of action theory, the ultimate goal is to give an account of "full-blooded agency," or agency "*par excellence*," which distinguishes between what merely happens to a person and what she does. A successful theory will be one that explains what "plays the role of the agent," as J. David Velleman (1992) puts it, or elucidates "the structure of a person's will," as Harry Frankfurt (1988) says, such that when this structure guides behavior, "the agent governs himself," in Michael Bratman's (2013) words. In other words, they are theories of autonomy.

*Email: skpaul@wisc.edu

According to Kantian approaches to autonomy, to give a theory of self-governance is at the same time to be engaged in doing moral philosophy. A Kantian agent is not fully autonomous unless he acts on maxims that satisfy the Categorical Imperative, and thus in conformity with and out of respect for the moral law. In contrast, the approaches to autonomy mentioned above deny that it is attainable only by those whose will is, or is aiming to be, morally good. These non-moral conceptions of autonomy allow that a perfectly self-governing Iago or Caligula is possible, in that the authority with which a given psychological structure speaks for the agent does not intrinsically depend upon a concern for moral reasons. Rather, the insight – the deep and important insight, in my view – is that there is an essential feature of agency *par excellence* that is shared by the sinner and the saint.

There is a second respect in which these contemporary accounts of autonomy differ from Kant's, and could perhaps be said to be more Aristotelian instead. This is in recognizing limits on the role reflective deliberation can or should play in responding to our reasons. On somewhat different grounds, Frankfurt, Bratman, and Richard Holton each claim that the autonomous agent will not consider in deliberation many courses of action which are in fact available to her. Indeed, the agent would be wrong to do so; to use John McDowell's apt term, certain potential options should be "silenced" for her (1979). On these views, the psychological structure that plays the role of the agent functions in part to filter out certain courses of action as inconsistent with one's plans, threatening to one's long-term goals, and at the limit, in conflict with one's entire practical identity. An essential aspect of self-governance is thus being such as not to have one deliberative thought too many.

My interest here is to ask, decades after Anscombe's prescription, what implications these morally neutral conceptions of autonomous agency have for ethics. In particular, what can we learn about the conditions of moral responsibility from our best current understanding of the psychology of self-governed action? I will argue that the proposed relationship between autonomy and deliberative silencing constitutes a pressure point for our thinking about agency and responsibility, forcing us to address the question of whether being autonomous is sufficient for being appropriately subject to blame, praise, and guilt. If we accept that it is (and I believe it is difficult to deny), then I will argue that these conceptions of autonomy have the consequence that we should be considered culpable for unwitting failures – things it simply never crossed one's mind that one ought to do, or ought not to do. Moreover, we should be considered culpable even when our failure to consider the morally correct option cannot be attributed to some prior choice of which it was reasonably foreseeable that such oversights would result. This verdict conflicts with a widespread assumption in moral philosophy that non-culpable ignorance is a general moral excuse (e.g. Rosen 2003). Some may therefore take this implication of contemporary action theory to be grounds to reject the theory, or the seemingly obvious connection between full-blooded agency and responsibility. Others may take it to be an illuminating result of following Anscombe's advice to do action theory before ethics. In either case, it will stand as a reminder that good intentions will not suffice to avoid the road to hell.

1. Control and the searchlight of conscious awareness

It is a bedrock principle of most theories of responsibility that ignorance is sometimes a moral excuse. This principle is traceable at least to Aristotle, who observed that pardon is bestowed on actions that are involuntary, and that "Those things, then, are thought involuntary, which take place under compulsion or owing to ignorance" (*Nicomachean Ethics*, Book III). The claim is not that we are never culpable for what we do as a result of ignorance; we do sometimes take an agent's ignorance of what she is doing, or failing to do, to be

culpable in its own right. If her ignorance is a result of being reckless or negligent in her obligations to be aware of morally relevant considerations, then she may be considered blameworthy for consequences that are themselves brought about unawares. A classic example is ignorance that is the result of a prior choice to become intoxicated, from which the likelihood of future negligence was reasonably foreseeable. But barring a violation on the basis of which the agent can be held responsible for her ignorance, the common thought is that we are not responsible for those features of our acts or omissions of which we were at no point aware. As Michael Zimmerman puts the claim, "all culpability can be traced to culpability that involves lack of ignorance" (1997, 418).

Deep questions immediately arise concerning what ignorance is, in the sense that bears on moral excuses. My interest here is in a particular understanding of ignorance that is connected to deliberation and choice, though I do not intend to suggest that this understanding is exhaustive of what is morally important about ignorance. On this conception of ignorance, which we might call "practical ignorance," the crucial distinction turns on what is accessible to conscious awareness, and thus poised to play a role in determining what the agent intentionally does. An agent might be practically ignorant of some consideration even if she subconsciously "knows" it to be true, if that consideration is not consciously accessible and thus unable to influence person-level deliberation. This type of ignorance can prevent an agent from having knowledge of the fact that a particular act or omission is a violation of some moral requirement even if she does not in general lack moral knowledge. The concept of practical ignorance enables us to formulate a more specific thesis concerning ignorance as a moral excuse: that conscious awareness at some point in time of at least some of the morally relevant features of an action or omission, or risk thereof, is required for moral responsibility. Following George Sher, I will call this thesis the "Searchlight View" (2009).

The Searchlight View is widely assumed, though not often explicitly defended, and is highly attractive for at least two reasons. The first is its apparent connection to what Thomas Nagel (1979) has labeled the Control Condition on moral responsibility: that one should not be morally assessed for what is due to factors beyond one's control. It is not a conceptual truth that being in control of some factor requires being consciously aware of that factor, but it is almost irresistible to think that the two are closely related. For example, Neil Levy takes as a premise in defending the Searchlight View that control requires knowing what one is doing, for "If I do not know either *that* I cause such changes [in the state of affairs], or *how* I alter the state of affairs, then I do not control it" (2005, 5, emphasis in original). The thought seems to be that merely being in a position to cause some outcome is not sufficient for being in control of it; rather, one must be in a position to cause or refrain from causing it *intentionally*. And the paradigmatic case of intentionally bringing about or allowing the occurrence of some state of affairs requires awareness of what one is doing: the agent must know what it is she aims to bring about. This suggests that an agent who lacks awareness that her action or omission has some morally relevant property P does not control whether she does something with property P, and therefore should not be morally accountable in that respect.

These reflections on the nature of control lead to the second reason the Searchlight View is attractive, which is its perceived connection to the expression of the will. The notion of the will can seem occult, but a relatively benign way to understand "willing" includes (though is perhaps not limited to) the making of a conscious decision. When we make a decision about what to do in the awareness that it will have a particular result, or that other alternatives are available, we effectively accept the anticipated results; they are included in the choice either as goal or foreseen side effect.[1] But in the absence of such awareness, the thought is that the will is not active either in choosing

or accepting that result. If willings are the proper object of moral assessment, the Searchlight View serves to capture a constraint on what it is to express the will through the making of a conscious choice.

Advances in social psychology and neuroscience have lately inspired challenges to the existence of the conscious will. The worry is that the phenomenology of conscious willing is an illusion: that actions are often initiated unconsciously and carried out automatically or guided by our implicit biases, that our situation rather than our character often explains what we do, and so forth (Wegner 2002; Bargh 2007; Bear and Bloom 2016). Part of my aim here is to raise a challenge to the importance of conscious agential control that sidesteps the need to argue on grounds of empirical research. I think it is our best philosophical theories of agency, rather than our best psychological experiments, that cast doubt on the Searchlight View.

2. Deliberation and the unthinkable

Whether or not particular acts of will essentially involve conscious awareness, a central insight of the work of Frankfurt, Bratman, and Holton is that autonomous self-governance over time is not merely a matter of a series of willings. On these views, to be self-governing is for one's psychology to have a certain structure, such that when this structure functions properly to guide thought and action, the agent governs. And although these views differ with respect to the details of this psychological structure, they agree that autonomy consists partly in being disposed such that some practical considerations never cross one's mind during deliberation. This is made fully explicit only by Holton, but I believe it follows also from the other two accounts if combined with a plausible view of the relationship between propositional attitudes and occurrent mental events.

The notion of the "unthinkable" is originally owed to Bernard Williams, but is developed by Frankfurt in a series of important papers exploring the relationship between personhood and the limits of the will (1988, 1998). His well-known claim is that having restrictions on one's choices, far from threatening one's freedom, is in fact crucial for one's identity. If all choices are on the table, he suggests, we become disoriented and uncertain, and the will is impaired. An autonomous will requires limitation in the form of volitional necessities stemming from what we care about and identify with. These volitional anchors render certain actions unthinkable for a person and thereby define the boundary of who she is.

Frankfurt generally does not intend for this term to be taken literally; the unthinkability he has in mind is most centrally an un-willability. Even if the agent does explicitly consider the option of, say, disowning his child, he will not be able to bring himself to do it. But it is a short step to suppose that sometimes, the structure of one's will imposes limits on deliberation by preventing the idea from even crossing one's mind. Frankfurt appeals at times to Bernard Williams's famous slogan "one thought too many" to describe the effect of caring, even surpassing Williams in arguing that when one's wife is drowning, "the strictly correct number of thoughts for [the husband] is zero" (2006, 36). That is, the husband should be able to act directly on his reason to save his wife without entertaining any thought or deliberative question whatever. Similarly, he claims that for the person who loves living, "[s]uspending his eagerness to live while wondering whether he has any other reason to go on would strike him as a clear case of having one thought too many" (1998, 172). For Frankfurt, then, the process of becoming an autonomous person capable of full-blooded action will involve the silencing of certain considerations or options in deliberation – the very structure that constitutes her identity will dispose her not even to think of these options as available.

Bratman's approach accords a central role to plans and policies rather than caring, but it has similar implications for conscious deliberation. On his view, the psychological structure whose guidance constitutes the agent's governance consists of a network of plan-states, and especially of higher-order self-governing plans concerning which considerations to treat as reason-giving in deliberation (1987, 2007). Plans are intentions "writ large," where intentions are understood as *sui generis* mental states on par with but not reducible to belief and desire. The key point here is that part of the functional role of intention on Bratman's view is to constrain deliberation to a tractably narrow set of options. Because we are creatures with finite cognitive capacities and limited time and opportunity for deliberation, we have practical need of such deliberative constraints; we would be paralyzed if we attempted to consider all of the practical options and arrive at a conclusion about what is best, full stop. Intentions therefore facilitate deliberation by providing a "filter of admissibility" for options (1987, 33). They also enable future planning by serving as stable anchors over time, tending to resist reconsideration in the absence of good reason to think one has made a mistake.

Like Frankfurt, Bratman is not specifically making a point about the thoughts that will be occurrent in deliberation. An intention will generate default pressure against treating an option as admissible even if that option does occur to the agent. But again, I think it will be a straightforward implication of his view that sometimes, the very plan structures that ground self-governance will cause the agent not even to consider that an option is available to her. Intentions could not play their functional role of streamlining deliberation and resisting reconsideration if the options that have been ruled in and out continue to be rehearsed in deliberation. An agent who is governed by a network of plans and policies will be disposed not even to think about options that conflict with those policies. In other words, a Bratmanian agent too is one whose structures of self-governance serve to prevent having one thought too many.

I suspect that this point sometimes goes unappreciated out of a tendency to idealize what deliberation is, conflating it with decision theory.[2] Even the most thorough episode of actual deliberation is not a matter of consulting one's entire repository of beliefs and desires and assigning expected utility to all available options. Human deliberation is far more constrained than this, and plausibly does not directly involve propositional attitudes at all. Rather, when I deliberate about what to do, a limited range of options occurs to me, and I decide between them (usually) by making judgments about the true and the good. These occurrent judgments are normally caused by my beliefs and desires, but they will not be a perfect reflection; some of my beliefs and desires will be left out, and my judgments will sometimes fail to accord with what I truly believe and desire. Once we recognize the distinction between propositional attitudes, which are diachronic and heavily dispositional states, and occurrent judgments, which are dated mental events, it is natural to understand the attitude of intention as a state that functions in part to constrain the occurrent thoughts we are disposed to undergo in deliberation.[3] If I have a standing policy of never taking the elevator, this will dispose me (particularly as time passes) not even to think of the possibility.

Holton (2009) insightfully emphasizes that certain types of intentions he terms "resolutions" are explicitly designed to function in this way, serving primarily to prevent certain practical options from presenting themselves in conscious thought. In addition to their role in streamlining deliberation, Holton accentuates the way in which intentions can enable us to overcome temporary shifts in our preferences and evaluative judgments due to temptation. On his view, a resolution in favor of a particular course of action succeeds in being stable in the face of the temptation to abandon that plan in part by inducing resistance in the moment to any direct thought of how one's reasons stand, akin to the way that

Jesuits attempt to resist thoughts of sin (124). Holton is not as directly concerned with giving an account of agential autonomy as Bratman or Frankfurt, and so may not seem to belong in the target group. But in fact, the very possibility of identifying an event as the "overcoming of temptation" depends upon supposing that the temptation-driven desire lacks agential authority while the overcoming-mechanism has it. Holton is therefore committed to viewing resolutions as wielding agential authority, and in this aspect his view is similar to Bratman's.

The general point is that on these conceptions of autonomy, agency *par excellence* is not merely a matter of controlling overt behavior. It will be part of the proper functioning of this network of intentions, resolutions, policies, and cares that certain considerations are simply silenced for the agent, such that they will tend not even to cross her mind when she is deliberating about what to do. And it is important to emphasize that it is no artifact of the approach that agential guidance is not limited to those things of which we are consciously aware. It is an explicit aim of these theories to avoid the temptation of explaining agency by positing a "homunculus" in the head that does the guiding, and the most tempting homunculus of all is consciousness. It can seem irresistible to suppose that I actively determine for myself only those things that I am aware of, or would become aware of if I attended to them. These theories of agency aim to avoid this mistake, holding that we are sometimes most active with respect to what we do not even consciously consider.

3. Unwitting culpability

The question now arises of what this means for our thinking about moral responsibility, in the sense of being appropriately subject to praise, blame, and guilt. What should we make of cases in which a set of self-governing commitments that are exemplary from the perspective of autonomy lead to morally bad results? The interesting cases will be those in which the agent fails to do something she ought to have done, or does something she ought not to have done, simply because the moral considerations in play never occurred to her as a result of her self-governing commitments. Such cases will be possible even if those self-governing commitments are morally permissible in their content, in that they display no ill will, criticizable evaluative judgments, or moral indifference. Nor need these lapses stem from general moral ignorance or a thoroughgoing incapacity to remember or attend to moral reasons. An agent who is generally capable of being sensitive to her moral reasons may nevertheless fail to consider them in a given case as a result of the proper functioning of the very structure that constitutes her as an autonomous agent.

As an example, consider Professor Plum. Plum is an autonomous, self-governing agent *par excellence.* He resembles Grahm in the epigraph from *Driftless*, but with respect to academia: he cares deeply about his academic pursuits and identifies with so caring. And he has formed a number of self-governing policies over the years concerning which considerations to treat as reasons. He modeled these after his professors in undergraduate and graduate school, because he values tradition and sees himself as tasked with perpetuating a venerable academic institution. In his view, part of the role of a professor is to be unconcerned with bourgeois social customs, and so his policies include giving significant weight to considerations of non-interference with academic freedom and little weight to the social dynamics of his workplace. Let us add that he has made a resolution to get a lot of writing done this semester, and so to avoid getting mixed up in department gossip.

The normal functioning of Plum's cares, policies, and resolutions in shaping what he considers in deliberation lead to the following two events. First, Plum is exposed to a bit of evidence that his colleague is behaving inappropriately to female students in the

department. It is far from conclusive evidence, since the situations he hears about are ambiguous, and it would require further investigation on his part to discover whether there is genuine wrongdoing (which there is). However, it simply never crosses Plum's mind that he could or should take action; his policy favoring non-interference with his colleagues' academic freedom and his resolution not to get involved in department gossip prevent it from even occurring to him. The second event is that after deliberating about how best to train his graduate students, he decides to require all of them to attend a conference that will be quite expensive for them (they will get no funding for it), and to be present at after-hours networking events that take place in bars. Plum conceives of this simply as the kind of good training he himself benefited from, and it never occurs to him that his policy might be seriously unjust to the students without outside financial support who cannot afford it and to the female students who tend to be treated very differently than male students at these after-hours events.

Plum is not suffering from general moral ignorance, and he does not harbor negative evaluative judgments concerning female or low-income students. If his practical thinking and action were not so heavily guided by his self-conception as a professor and the commitments and policies that structure this practical identity, he is the kind of person that would be more attuned to the opportunities to correct social injustice that he is presented with. Indeed, he is sensitive to this kind of moral reason in settings other than his professional life; his behavior is not straightforwardly explicable by a general lack of concern for doing right. Most importantly, Plum's intentions are good, at least in the sense that the values that structure them really are worthwhile considerations, and in that there is nothing explicitly criticizable in their content. He is simply deaf to some of the moral features of his situation.

The question is whether agents like Plum are morally blameworthy for what they do, or fail to do, when the fact that they should act otherwise simply never occurred to them as a result of their self-governing commitments. I think cases such as these pose a genuine dilemma, and I do not aim to argue here that there is only one reasonable response. What I do wish to argue is that if we accept a quite modest principle linking agentive activity with responsibility, it follows from the "silencing" views of autonomy that agents like Plum ought to be held responsible, on pain of those theories conceding that they are not really accounts of agency at all. Some may consider this result to be a *reductio* of the approach, in that it conflicts not only with the Searchlight View about moral responsibility but also with most versions of its major competitor, Attributionism. On the other hand, others who are friendly to the approach and thus accept the consequence will be led to a fairly revisionary view of the conditions of moral responsibility. In what follows, I will first argue for the implication, and then reflect on the case for and against accepting it. As I think some ambivalence is genuinely warranted here, I will ultimately leave it to the reader to decide which horn to grasp.

The reason it follows from the Silencing View of autonomy that people like Plum should be blamed for failing to respond to considerations that never occurred to them is simple: these events are direct manifestations of their agency. The following principle states a modest connection between agential activity and responsibility, and strikes me as true: if we are morally responsible for anything in our lives, we are responsible for those instances in which we are most active in determining ourselves. This principle is consistent with concluding that we are not morally responsible for anything, although I will assume here that we are. It is also consistent with supposing that responsibility extends even further to events with respect to which we are passive, but that would be a stronger thesis than is needed here. The weaker principle is simply meant to capture the widely-held

thought that the most appropriate candidates for praise and blame are those things that are up to us. And this is just what the views in question are theories of: they articulate what it is to be active as opposed to be passive in our lives. A central insight is that agential control is not limited to conscious control, and the self is not something that is simply passively inherited. But if this is right, then there seems to be no room to withhold responsibility from Plum; his thoughtlessness was the direct manifestation of his capacity to be a self-governing and autonomous agent, functioning precisely as it was supposed to.[4]

One might worry that this argument proves too much, since we are not responsible for all of the properties of events that are exercises of full-blooded agency under some description. If I intentionally put a substance in your coffee that I have every reason to believe is sugar, the fact that the event was intentional under one description should not entail that I am blameworthy for poisoning you with what was in fact arsenic.[5] It is tempting to conclude that this is because the poisoning was unintentional, and infer that because the silencing of certain considerations in deliberation is also unintentional, it is also excluded from the scope of responsibility. What I am suggesting, however, is that the status of an event as unintentional does not necessarily bar it from being a manifestation of agency, and therefore from moral assessment. On the views in question, what intentional actions and silencing have in common is that realizing the content of the thinker's intentions and structuring her deliberation are both defining goals of the psychological system that constitutes her as an agent. This system was guiding in its characteristic way when it led Plum to have one thought too few; that the relevant considerations were filtered out of deliberation was not an accident or a byproduct of self-governance, but rather an essential part of it. Bringing about unforeseen side effects, in contrast, is no part of the system's defining goal and so not a direct manifestation of agency.

It is open to the Silencing View to attempt to avoid the conclusion that Plum is responsible for his unwitting violations by maintaining that the notions of activity, control, and guidance it is elucidating are different in kind from those appealed to in the above principle, and that they are not intended to have any direct implications for deep responsibility. To employ a distinction of Gary Watson's (1996), perhaps these views are relevant only to attributability and not to genuine accountability. This would clearly be false of Frankfurt's work, which is explicitly addressed to questions of moral responsibility, but I think it is consistent with the dominant focus of Bratman's and Holton's contributions (see Holton 2010). Nevertheless, it would be disappointing to discover that the research conducted in the philosophy of action for the last few decades simply has nothing to do with the question of what we are accountable for. On pain of being vulnerable to the accusation that we action theorists have entirely changed the subject, this option ought to be avoided if possible.

Given this, let us examine the considerations bearing on whether to accept that we are blameworthy for such unwitting failures. The first thing to note is that this consequence conflicts with the verdict of both the Searchlight View and most existing versions of its major alternative, the Attributionist approach. The Searchlight View will not hold him accountable because there was no point in time at which Plum chose a course of action in the awareness of its negative moral implications, or of which such implications were reasonably foreseeable. The fact that the relevant considerations never entered his mind during deliberation was not the product of a process of which he was directly aware or about which he made a conscious choice. True, there were points in time when Plum consciously reflected on the self-governing policies and resolutions that contributed to these moments of moral deafness.[6] Only an antecedent allegiance to the Searchlight View would justify us in insisting that these past events are the locus of blame, however. He adopted his self-governing policies concerning the limits of his role and responsibilities

as a professor many years ago, and although they could have been more morally admirable, they were not impermissible in their content. Nor was it reasonable to expect him to foresee the impact they would have on his far future deliberation (as Vargas 2005 convincingly argues). Even more plausibly, when one comes to care about a person, career, or cause, the fact that it may lead to future moral deafness is generally either unforeseeable or irrelevant – we are not required to avoid caring about things just because it might have bad consequences down the road.

The first question is thus whether we have stronger reason to adhere to the Searchlight View than to embrace the consequences of these views of autonomous agency. The intuition is powerful that if Plum was never consciously aware of the moral wrongness of his choices, or risk thereof, then his infractions were too akin to inadvertence to merit blame. If his failure to notice the moral downside of his behavior did not express reprehensible evaluative judgments about the value of other people, how is this relevantly different from inadvertently becoming distracted or, to use an example of Frankfurt's, failing to suppress a burp (2008)? But I am effectively suggesting that the Silencing View of agency offers a ready answer to this question: unwitting violations are culpable when they are the direct result of agential activity, whereas genuine inadvertence is passive – it is something that happens to you. Viewed in this light, the heart of the disagreement is over whether anything could truly count as active self-determination if conscious awareness is not involved.

But this is primarily a question for action theory, and so I suggest that the burden is now on the Searchlight View to explain why not; mere intuitions to the contrary will not suffice. Again, the suggestion cannot simply be that conscious awareness is like a little person in the head that does the deciding. What we have seen is that even if it is granted that intentional *action* necessarily involves awareness of what one is doing, this would not suffice to vindicate the Searchlight View of responsibility. The assumption that we can only be held responsible for our intentional actions and not unwitting omissions or unforeseen side effects would beg the question in this context.

Levy (2013) has commendably taken up this burden on behalf of the Searchlight View, arguing that awareness is essential because it is a necessary condition of expressing one's evaluative stance on the world through action. The first premise in his argument is that an agent's values consist only in attitudes of which she is access-conscious. In support of this claim, Levy proposes that only conscious attitudes generate the dispositional stereotype that is associated with or constitutive of valuing, because unconscious attitudes tend not to be informationally integrated or rule-based and thus fail to realize the normative aspects of that dispositional profile (218–221). The second step in the argument asserts that because deliberation is itself a conscious process, only conscious beliefs are available to form the content of our reasons for action (222–224). The conclusion is that only actions settled on consciously in deliberation and caused by reasons of which the agent is aware are genuine expressions of the quality of that agent's will, and thus the only events for which we are responsible (227). Awareness plays a double role, underwriting the status of a set of attitudes as the agent's values and enabling the expression of those values through action.

For present purposes, let us simply grant Levy's premises; according to the theories of agency under discussion, the conclusion does not follow. In particular, let us grant that our values, understood as the ends we treat as reason-giving, must feature consciously in deliberation in order to serve as the content of our reasons for acting. It does not follow that the will is only expressed in the content of the reasons for which we act, or that only an agent's evaluative outlook is subject to moral assessment. Precisely what I have aimed to highlight here is the way in which the will may be directly implicated in what we do not do, or do

unwittingly. It might be that the cares, policies, and resolutions that constitute the will must themselves be accessible to consciousness (though I am not convinced that this is so, especially in the case of cares), but their agential manifestations will not always be conscious in nature. In the absence of further argument, then, the Searchlight View is left in the position of begging the question against the Silencing View of agency, and so cannot serve as the basis for rejecting the implications of the Silencing View.

So much the worse for the Searchlight View, we might think. But perhaps surprisingly, most extant versions of the major alternative, the Attributionist View, will not deliver a different verdict on Plum. Attributionist views deny that responsibility must be grounded in a conscious choice that was made at some point, holding instead that an agent is responsible for an event or state of affairs just in case it can correctly be attributed to her as expressive of something relevant about her psychology or character. However, they tend to hold that attitudes, omissions, failures to remember, and the like are blameworthy only insofar as they specifically reflect something morally problematic about the agent's evaluative judgments or a lack of good will. On Nomy Arpaly's (2003) view, for instance, the failure to do what is morally required is blameworthy only if it reflects a lack of concern for one's moral reasons. Angela Smith (2005, 2008) has argued that one is morally responsible for some event if that event is a manifestation of the fact that one's rational judgments are in violation of moral standards. And T. M. Scanlon has similarly claimed that an agent is blameworthy for those things that express or reveal the presence of attitudes toward others that impair the relations others can have with her (1998, 2008).

Attributionist views are in a good position to hold people responsible for what they fail to notice, remember, or think of. The plausible thought is that such lapses can express the presence of morally objectionable attitudes or the absence of moral concern even if this expression is involuntary. But even granting that this is correct, the problem is that there is no firm ground on which to insist that Plum must either have morally objectionable evaluative attitudes or lack proper moral concern. To assume that this is so is essentially to presuppose that if he had had the proper concern, or had lacked some of his actual evaluative attitudes, the relevant considerations would have occurred to him. From the perspective of moral psychology, this presupposition would be bizarrely ad hoc: why think that even a fitting amount of moral concern *necessarily* overcomes the proper functioning of agential silencing mechanisms? This is not merely an unreasonably high bar to set, but also metaphysically inexplicable. And it is explanatorily unnecessary to posit a deficient desire to do what is right; the more obvious explanation appeals to the strength of the silencing mechanism. Plausibly, it must not be the case that silencing always or even frequently overrides moral sensitivity, but an agent's cares, policies, and resolutions may prevent her from putting two and two together in a particular case even if her desire to do what is right generally does guide her practical thought. By hypothesis, the self-governing commitments that led to Plum's lapses are not themselves reflective of negative evaluative judgments about women or those with meager financial resources. To be sure, he could have made moral sensitivity more of a priority. But I take the lesson of Susan Wolf's (1982) repudiation of moral sainthood as a personal ideal that Plum's emphasis on non-moral values is permissible.

Sher is unique (to my knowledge) in defending a morally neutral Attributionist account, holding that the agent should be identified with "the collection of physical and psychological states whose elements interact to sustain his characteristic patterns of conscious and rational activity," and held responsible for the upshots of that structure with or without awareness (2009, 124). He does not aim to spell out in detail what this structure consists in, but I suggest that this is precisely what contemporary action theory has been engaged

in. There is therefore precedent for an Attributionist view that takes Plum to be blameworthy for his unwitting failures. It is important to emphasize, however, that Sher himself takes this upshot to be a radical reconceptualization of the epistemic conditions on moral responsibility. If our best theories of autonomous agency vindicate this result, we will have learned something significant indeed by doing action theory before ethics.

I will conclude this section by reflecting on whether there are independent reasons to find this result acceptable. In favor of holding Plum responsible, it is common to say that he should have known. While attractive, this claim is highly puzzling. In some cases, it is merely a way of expressing the idea that Plum has conducted himself in a way that had negative consequences, and that it would have been better if he had known. But this is something on which all sides can agree. In the law, it is understood as the claim that another person – a reasonable person – would have known. It is unclear how the epistemic status of a different person can ground Plum's responsibility, however. Plum himself was unaware of what he was doing, and so his decisions themselves did not entail an acceptance of the problematic consequences even if those of his more reasonable counterpart would have. Moreover, we cannot simply infer from the existence of this legal standard that it captures anything illuminating about the conditions of moral responsibility, for the law is concerned with a variety of factors other than enforcing morality.

We might attempt to forge a connection between reasonable persons and actual persons by claiming that in every case where a reasonable person would have known, there was something the actual person should have done in order to have known. This is the right thing to say in tracing cases, as when the ignorance is a result of the agent intentionally having become intoxicated. But in most instances, there will have been no explicit choice to render oneself incapable of attending to one's moral reasons. We might say instead that the reasonable person would have known because there was evidence available from which he would have reasoned to the correct conclusion, and the actual agent could have done the same. But while the accessibility of relevant evidence is plausible as a necessary condition of responsibility, I do not think it is sufficient. As Rosen (2003) points out, it might be true that an agent could easily have arrived at the correct conclusion if he had reflected, but he might also have had no special reason to suppose that further reflection was called for. If we are never permitted to refrain from reflecting on evidence that is accessible to us, then it is clear that the silencing effect of self-governing commitments will be morally problematic – but surely such a moral obligation is too demanding for cognitively-limited agents like us.

The most plausible justification for holding Plum responsible is to claim that he ought to have been more attentive and open-minded. If he were such as to consider more and various options in deliberation, he would have come to a different and better conclusion. Barbara Herman has argued, for instance, that maxims of sufficient means must be adopted with respect to ends that are morally required (1993, 99–101).[7] In some cases, the required means will involve ensuring that one remembers or is otherwise in a position to fulfill one's obligations. If this is right, then perhaps the correct criticism to make of Plum is that he failed in his duty to do whatever it takes to ensure that he does not blank on important considerations.

But this diagnosis would have radical implications for anyone sufficiently committed to being morally good: we must be wary of becoming autonomous, self-governing agents, assuming that the Silencing View is correct about what this amounts to. Doing whatever one can to be aware of all the potentially relevant considerations and to weigh them properly in deliberation is potentially in tension with having general policies, resolutions, or cares that may prevent one from noticing such considerations in a given case. Perhaps

Plum ought not to have cared so deeply for carrying on tradition and focusing on his research, or to have formed policies leading him to ignore the interpersonal dynamics of his profession. The values to which he committed were in general permissible; this is what makes him very different from Iago, whose self-governing commitments are explicitly evil. But on one reading, the lesson he teaches is that in pursuing autonomy, it does not suffice from a moral point of view to act only on good intentions.

As I said at the outset, I see no easy way of answering the question of whether to blame Plum. I will leave it to the reader to decide whether the argument is an instance of Modus Ponens or Modus Tollens.[8]

4. Self-governance, morality, and rationality

I will close by reflecting briefly on the significance of these observations for our thinking about the value of autonomy. Others such as Arpaly have argued that autonomy is not a necessary condition for direct moral praise- and blameworthiness. The question of this paper has been whether it is even sufficient. If it is not, I think this would show that contemporary action theory has gone significantly astray, to the point of changing the subject. If it is, then it turns out that cultivating a set of self-governing commitments that are in themselves unobjectionable can lead to blameworthy deliberative inattentiveness down the road. Enhancing agential control comes at the price of curtailing sensitivity to situations that were unanticipated in those commitments.

This is not to say that we should not strive for self-governance, or should avoid developing a wholehearted and resolute character; there are permissible personal ideals other than moral sainthood. However, I do think that the potential tradeoff between achieving a resolute, self-governed, volitionally-anchored will and being sufficiently sensitive and open-minded to one's moral reasons may have important consequences. For instance, I think it is a basis for concern about the project of grounding the normativity of rationality in the value of self-governance, as Bratman has recently advocated in an important and fascinating series of papers (2009, 2012). Bratman's proposal is that we have a reason to govern our own lives, and that this global reason grounds a local reason to be self-governing at each particular opportunity in which self-governance is possible. He further argues that satisfying the rational norms on intention – intention consistency, means-end coherence, and an intention-stability norm he calls (D) – is a necessary constitutive means of being self-governing. If we do not satisfy these rational norms, our practical standpoints will be indeterminate in a way that blocks the possibility of self-governance. We therefore have reason to be rational in each case in which self-governance is possible. The claim is not that this reason is overriding in every case, or even especially strong; it might be that we have better reason in some instances not to be rational. But the desired conclusion is that absent some deep psychological defect preventing the possibility of self-governance, there is always some reason to be practically rational.

I have elsewhere voiced doubts about whether the average concern for self-governance is sufficient to generate a reason to be self-governing at each opportunity (Paul 2014). Most of us also value spontaneity and whimsy, such that a life of perfect self-governance would be a failure by our own lights. But I think it is even more worrisome for this strategy that we might have moral reason to curtail self-governance. An agent who actually availed herself of every opportunity to be self-governing is much more likely to find herself in the situation of failing to think of morally relevant considerations in a given episode of deliberation. The claim is not that being self-governing necessarily leads to moral violations; for all I have argued, it is possible to be autonomous in a morally laudable way. But for the worry to

arise, it suffices that in pursuing a self-governed life, we can rather easily end up in a state like Plum's without making any particular choice with recognizably immoral content.

For an agent in this condition, her reason to be self-governing competes with her reason to do what it takes to be moral, as well as any other conflicting personal ideals. This conflict will be especially vivid if there is no way for her to improve morally without temporarily undermining self-governance, as will at least sometimes be the case. Now, it is a complicated matter to understand how general reasons – reason generally to be self-governing, reason generally to be moral – compete to issue reasons for particular actions that would promote one value but not the other. It is plausible to suppose that in some cases, however, one's reason to be self-governing will be completely defeated by the other reasons in play. That is, in a given instance, one may have no normative reason at all to pursue self-governance even if it is perfectly possible to do so.

In contrast, the normative force of the requirements of rationality does not seem subject to being canceled out by other substantive concerns, or to wax and wane depending on what else is at stake. If rationality is a good, it is generally thought to be a good that is independent of one's substantive ends. I do not pretend to have argued for this, and a more thorough defense of these claims is beyond the scope of this paper. But I tentatively conclude that grounding the normativity of rationality in the value of self-governance leaves rationality on shaky ground. Exemplary self-governance can directly pave the road to hell in a way that exemplary rationality should not.

Acknowledgements

Many thanks to Matt Smith and Ulrike Heuer for their comments and for their work in editing this issue. I am grateful to Michael Bratman, Richard Holton, Jennifer Morton, Kenny Walden, and Ben Schwan for helpful discussions and comments, as well as to all the participants of the "Moral Significance of Intention" conference at the University of Leeds in June 2014 and to audiences at the University of Pittsburgh, Ryerson University, and UC-Irvine.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes

1. Castañeda (1979) eloquently articulates the phenomenon of the "acceptance" of consequences in the context of practical deliberation:

> An action that one ponders and places as a side action in a plan leading to a goal action, is an action that one . . . accepts in spite of how painful it is, in order to attain that goal. This deliberate toleration is of the same family as the acceptance we call intending.

2. Though of course, not by everyone; for instance, the point has been made very clearly by Matthew Smith (2010) and Arpaly and Schroeder (2012).
3. As Ben Schwan argues in "The Reasons Intentions Give," unpublished manuscript.
4. There are of course numerous senses of responsibility afloat in the philosophical literature. My hope here is to sidestep the complexity of this debate by focusing on a broad and untheorized notion of responsibility, stipulating only that it must be one that grounds praise, blame, and guilt. I think an upshot of the argument is that the burden is on those who wish to deny that Plum is subject only to a weaker or superficial sense of responsibility to explain why, given that there is nothing lacking in his case that is obviously essential to his agency.
5. Thanks to James Shaw for raising this objection.

6. It is this feature of the view that make it less obviously subject to Susan Wolf's objections to "Real Self" accounts of responsibility.
7. Thanks to Claudia Card for pointing me to this.
8. Of course, nothing I have said here rules out the possibility of supplementing the Silencing View of autonomy with further conditions designed to bring its implications in line with what we independently take to be true of moral responsibility. For instance, one could add a "searchlight" condition on agential guidance, excluding any process of which the agent is unaware from having the status of self-governance. But many potential modifications, including this one, will be in tension with the foundational commitments of this approach to understanding agency and so will fall under the "Modus Tollens" heading.

Notes on contributor

Sarah K. Paul is an Associate Professor at the University of Wisconsin-Madison, and has held visiting positions at Bowdoin College, MIT, and NYU-Abu Dhabi.

References

Anscombe, G. E. M. 1958. "Modern Moral Philosophy." *Philosophy* 33 (124): 1–19.
Aristotle. *Nicomachean Ethics*. Translated by W. D. Ross.
Arpaly, Nomy. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.
Arpaly, Nomy, and Timothy Schroeder. 2012. "Deliberation and Acting for Reasons." *Philosophical Review* 121 (2): 209–239.
Bargh, J. A. 2007. *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes*. New York, NY: Psychology Press.
Bear, Adam, and Paul Bloom. 2016. "A Simple Task Uncovers a Postdictive Illusion of Choice." *Psychological Science* 27 (6): 914–922.
Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Stanford, CA: CSLI Press.
Bratman, Michael. 2007. *Structures of Agency*. Oxford: Oxford University Press.
Bratman, Michael E. 2009. "Intention, Practical Rationality, and Self-governance." *Ethics* 119 (3): 411–443.
Bratman, Michael. 2012. "Time, Rationality, and Self-governance." *Philosophical Perspectives* 22 (1): 73–88.
Bratman, Michael. 2013. "The Fecundity of Planning Agency." In *Oxford Studies in Agency and Responsibility*, Vol. 1, edited by David Shoemaker, 47–69. Oxford: Oxford University Press.
Castañeda, Hector-Neri. 1979. "Intensionality and Identity in Human Action and Philosophical Method." *Nous* 13 (2): 235–260.
Frankfurt, Harry G. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press.
Frankfurt, Harry G. 1998. *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.
Frankfurt, Harry G. 2006. *Reasons of Love*. Princeton: Princeton University Press.
Frankfurt, Harry G. 2008. "Inadvertence and Responsibility." *The Amherst Lecture in Philosophy* 3: 1–15. http://www.amherstlecture.org/frankfurt2008/.
Herman, Barbara. 1993. *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.
Holton, Richard. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.
Holton, Richard. 2010. "Disentangling the Will." In *Free Will and Consciousness: How Might They Work?* edited by Al Mele, Kathleen Vohs, and Roy Baumeister, 82–100. New York: Oxford University Press.
Levy, Neil. 2005. "The Good, the Bad, and the Blameworthy." *Journal of Ethics and Social Philosophy* 1 (2): 1–16.
Levy, Neil. 2013. "The Importance of Awareness." *Australasian Journal of Philosophy* 91 (2): 211–229.
McDowell, John. 1979. "Virtue and Reason." *The Monist* 62 (3): 331–350.
Nagel, Thomas. 1979. *Mortal Questions*. Cambridge: Cambridge University Press.
Paul, Sarah K. 2014. "Diachronic Incontinence Is a Problem in Moral Philosophy." *Inquiry* 57 (3): 337–355.
Rhodes, David. 2008. *Driftless*. Minneapolis, MN: Milkweed Editions.
Rosen, Gideon. 2003. "Culpability and Ignorance." *Proceedings of the Aristotelian Society (Hardback)* 103: 61–84.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Scanlon, T. M. 2008. *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge, MA: Harvard University Press.
Sher, George. 2009. *Who Knew? Responsibility Without Awareness*. Oxford: Oxford University Press.
Smith, Angela M. 2005. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115 (2): 236–271.
Smith, Angela M. 2008. "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138 (3): 367–392.
Smith, Matthew Noah. 2010. "Practical Imagination and Its Limits." *Philosophers' Imprint* 10 (3): 1–20.
Vargas, Manuel. 2005. "The Trouble with Tracing." *Midwest Studies in Philosophy* 29 (1): 269–291.
Velleman, J. David. 1992. "What Happens When Someone Acts?" *Mind* 101 (403): 461–481.
Watson, Gary. 1996. "Two Faces of Responsibility." *Philosophical Topics* 24 (2): 227–248.
Wegner, Daniel. 2002. *The Illusion of Conscious Will*. Cambridge: The MIT Press.
Wolf, Susan. 1982. "Moral Saints." *Journal of Philosophy* 79 (8): 419–439.
Zimmerman, Michael J. 1997. "Moral Responsibility and Ignorance." *Ethics* 107 (3): 410–426.