WHO KNEW?

Responsibility without Awareness

GEORGE SHER

OXFORD UNIVERSITY PRESS
2009
THE RESPONSIBLE SELF

How, if Alessandra, Julian, and the others have not consciously chosen either to act wrongly or to risk failing to realize that they are doing so, can the wrongness of their acts be attributable to them in a strong enough sense to render them responsible? The answer, I have suggested, is that even if each agent's failure to realize what he should is neither an action nor the foreseeable result of one, it may still have originated in him in the sense of resulting from the interaction of the very attitudes, dispositions, and traits that make him the person he is. To complete my defense of PEC, I must now argue for the conception of the responsible self that underlies this suggestion.

I

Even if our choices are not always causally determined, our failures to realize and remember things surely are. Thus, it is hardly controversial to say that whenever someone fails to realize what he should realize, or forgets what he should remember, his cognitive lapse can be traced to some combination of elements of his psychology and/or physiology. The harder question, though, is why this fact should have any bearing on the agent's responsibility. Why should our ability to trace a person's
cognitive lapse to the effects of certain subpersonal states have anything to do with our reactions to him?

To answer this question, I will have to show that a person's causally effective subpersonal states are so closely related to him that it is reasonable to view the cognitive failures to which they give rise—and so, by extension, the ensuing wrong or foolish acts—as originating in him. I will have to show, in other words, that the relevant subpersonal states are, in some suitable sense, among the person's constitutive features. The task of showing this is complicated by the fact that terms like "constitutive feature" and "feature that makes someone the person he is" are multiply ambiguous, and can be used to designate anything from a feature that is part of someone's metaphysical essence to one that is central to his self-image. However, because our guiding aim is to understand the conditions under which agents can be held responsible for what they do, the obvious way to sidestep these ambiguities is to allow the relevant conception of the self to be shaped by intellectual pressures that originate in the concept of responsibility itself. As we will see, these intellectual pressures are generated mainly by facts we have already encountered. They arise because only beings that are capable of acting for reasons can qualify as responsible, because each reason-responder is both embodied and endowed with a subjectivity that is associated with a unique perspective, and because those who hold agents responsible typically do not share either their subjectivity or the perspectives they occupied when they acted.

How, in light of such facts, should we think of the responsible self? To get an initial sense of the possibilities, let us begin by considering two extreme positions, one maximalist and the other minimalistic. The maximalist position identifies the responsible self with the whole human organism. It denies that any of an agent's physical or psychological features are any more constitutive of him than any others, and so insists that every (skin-in) fact about him, from his height or his white blood count through his most deeply held values, is equally part of what makes him the person he is.1 By contrast, the minimalist position takes its cue from the fact that many types of entity—stones, animals, and infants, for example—cannot be responsible for anything at all. Because responsibility evidently requires features that such entities lack, the minimalist reasons that any adequate theory of the responsible self must focus exclusively on these. This leads him to abstract away from all merely physical aspects of the person, and to view responsible selves as constituted entirely by characteristics such as rationality or conscious will.

Although both extreme positions embody important insights, I want to argue that neither is adequate. Where the maximalist position is concerned, the problem is obvious: although each responsible agent is of course a human being, and so is located in the natural world and is subject to its causal laws, the vast majority of each human's physical features (and perhaps also many of his psychological features) have no obvious connection to any of his beliefs about himself or the world, his judgments about what he has reason to believe or do, or his actual decisions, actions, and omissions. When a given physical or psychological feature is this dissociated from all the aspects of its possessor's life which alone make questions of responsibility meaningful, there is simply no reason to view it either as any part of what makes him a responsible agent or, a fortiori, as any part of what makes him the particular responsible agent he is.

Because an indiscriminately inclusive approach would draw the boundaries of the responsible self far too broadly, we must draw back from the maximalist position. But how far back should we draw? Given the wide consensus that any being that entirely lacked consciousness, or that sometimes or always had it but was systematically insensitive to theoretical or practical reasons, would not qualify as a responsible agent,2 there is an obvious case for taking consciousness and reason-responsiveness to fall within the boundaries of the responsible self.

---

1. The importance of the physical boundaries of the human organism has been challenged both on the grounds that the contents of beliefs are partially determined by external events and on the grounds that external events play an active role in driving cognitive processes; see, respectively, Tyler Burge, "Individualism and the Mental," *Midwest Studies in Philosophy* 4 (1979), 73–122, and Andy Clark and David Chalmers, "The Extended Mind," *Analysis*, 58 (1998), 10–23. If either challenge can be sustained, then there is room for a supermaximalist position according to which the responsible self is constituted by all of the agent's skin-in features and more.

2. For example, Thomas Scanlon—no friend of the searchlight view—has observed that "it is crucial to a creature's being a rational creature that conscious judgment is one factor affecting its behavior" (Thomas Scanlon, *What We Owe to Each Other* [Cambridge, MA: Harvard University Press, 1998], 282).
harder question, though, is which other features of human beings fall within those boundaries. If we answer "none," and abstract away from everything about the responsible agent except his consciousness or reason-responsiveness, then we will have arrived at the minimalist position.

Because so many reason-based choices are conscious, it may at first seem that any minimalist must take responsible agents to be constituted by both consciousness and reason-responsiveness. However, the issue is complicated by the fact that consciousness and reason-responsiveness can themselves come apart. It is quite possible to be a conscious chooser without being reason-responsive; and, as we saw in chapter 4, it is also possible to respond to reasons at a nonconscious level. Because consciousness and reason-responsiveness need not coincide, there is room for versions of minimalism which take responsible agents to be constituted exclusively by their capacity for conscious choice, exclusively by their reason-responsiveness, or by both together.

The distinction between accounts that represent responsible selves as conscious choosers and those that represent them as reason-responders is far from trivial. It is precisely this distinction that underlies the contrast between the variant of the searchlight view that Neil Levy labels "volitionism"—the view that "an agent is responsible for something (an act, omission, attitude, and so on) just in case that agent has—directly or indirectly—chosen that thing"—and the influential opposing view, espoused by Thomas Scanlon and Angela Smith and often labeled "attributionism," which asserts that agents are responsible for all of the actions, beliefs, and attitudes, conscious or not, that reflect their judgments about what they have reason to do, believe, or feel. Because volitionism abstracts away from everything except an agent's conscious choices, it in effect identifies the responsible self with, or in terms of, the first-person perspective from which he views the world. By contrast, because attributionism abstracts away from everything except an agent's judgments about reasons, it in effect identifies responsible selves with their capacity to reach such judgments.

Given the intimate connections between responsibility on the one hand and consciousness and reason-responsiveness on the other, both minimalist views clearly satisfy the requirement that responsible selves be constituted exclusively by features that have some important bearing on their responsibility. However, importantly, it is also possible to satisfy this requirement in ways that are less ruthlessly minimal. We will also satisfy it if we take each responsible self to be constituted not merely by his consciousness or reason-responsiveness, but also by certain physical or mental features to which consciousness and reason-responsiveness are themselves intimately linked. By thus working outward from these core features, we can identify a whole array of conceptions that stand between the maximalist and minimalist extremes. Although many of these intermediate possibilities need not concern us, there is one addition to the minimalist view that I think is definitely required. In what follows, I will first propose, and then defend, an account that incorporates that addition.

II

Put baldly and without defense, my basic idea is that a responsible agent is best identified not only with his subjectivity or rationality but also with their causes. Put a bit more specifically, what I will argue is that we should think of each responsible agent not merely as a conscious center of will, nor yet as an entity that attempts to evaluate its practical and theoretical reasons and to base its beliefs and actions on them, but rather as an enduring causal structure whose elements interact in ways that give rise to these responsibility-related activities. As a corollary, when I say that a given psychological or physiological state is among the constitutive features of a particular self, or that it is part of what makes him the person he is, what I will mean is simply that it is among the elements of the system whose causal interactions determine the contents of the conscious thoughts and deliberative activities in whose absence he would not qualify as responsible at all.
As so understood, each responsible agent’s constitutive features will be quite extensive. It is a commonplace that each person’s theoretical and practical decisions are influenced by factors such as his background beliefs, his moral commitments, his views about what is good and valuable, and what he notices and finds salient. His decisions are influenced, as well, by his degree of optimism or pessimism, his attitude toward risk, and many other facets of his emotional makeup. Hence, by my account, all such features of an agent will qualify as constitutive. In addition, as long as they remain compatible with the general framework of folk psychology within which the notions of agency, reasons, and responsibility are embedded, we can expect that many of the factors that explain an agent’s thoughts and actions at other, deeper levels—the relevant neurophysiological mechanisms, for example, or the functionally defined constructs that populate the flow charts of cognitive psychologists—will qualify as constitutive too. As I hope is obvious, these claims are meant to imply not that we cannot classify someone as a responsible self without being able to identify the relevant causal structures, but only that in so classifying him, we assume that such structures do in fact exist.

Because my proposal identifies each responsible agent with whatever psychological and physical structures sustain his normal patterns of intellectual functioning, it is clearly a variant of the familiar view that the identity of each such agent is determined by his character. It is, however, quite different from the familiar versions of this view that understand character exclusively in terms of behavioral or affective dispositions, deeply held value-commitments, or the relations between the agent’s higher- and lower-order attitudes. What sets my approach to character apart from these—and, not coincidentally, what promises to enable it to play the role that PEC assigns it—is of course its frankly causal nature.

There is, however, a question about its ability to fulfill this promise. Put most simply, the difficulty is that if we identify a responsible agent’s constitutive features by working backward from his reason-responsive activities to their causes, then it is hard to see how we can take those same features to give rise to acts or omissions that are not reason-responsive. If what renders an element of someone’s psychology constitutive is precisely its role in sustaining his rationality-related activities, then how can any feature by which he is constituted prevent him from engaging effectively in such activities? And, if this is not possible, then how can unwitting wrongdoers like Alessandra, Julian, and Scout be linked to their wrong acts by the fact that their failures to appreciate the acts’ wrongness were caused by their constitutive features?

To answer these questions, and thus complete my sketch of the view that I am proposing, I want to make two points, one conceptual and one broadly empirical. The conceptual point is simply that there is no contradiction in the claim that a feature that is constitutive of an agent in the specified sense can sometimes prevent him from responding to reasons. As long as the effects of any given element of an organized psychological system depend both on that element’s interplay with the system’s other elements and on many external factors—and this, of course, is always the case—it will be quite possible for a state that generally contributes to its possessor’s reason-responsiveness to interfere with it on a given occasion. We are all familiar with many normally but not invariably reliable systems—cars whose sparkplugs occasionally do not fire, for example, and computers that occasionally freeze—whose internal organization allows us to understand both why they are generally reliable and why they sometimes malfunction. Thus, if the current objection is to have bite, it will have to be because the aspects of the relevant agents’ psychologies that explain their cognitive lapses do not in fact fit this description.

But why, exactly, should we suppose this? Why, when Alessandra forgets about Sheba because she is distracted by the dispute at the school, should we suppose that the aspects of her psychology that account for her distraction—her concern for her children, for example, or her tendency to focus intensely on whatever issue is at hand—are anything but consistent contributors to the way she characteristically approaches practical and theoretical problems? When Ryland fails to realize that the members of her audience will find her anecdote offensive, why view the self-absorption that explains her failure as anything other than an unremarkable (if sometimes distorting) member of the vast collection of attitudes whose interaction determines the usual shape of her judgments about reasons? Even when Joliet panics and Father Poteet misjudges the available road space, why exclude either her emotional makeup or the speed with which he processes visual cues from the factors that play a role in determining the contours of their respective judgments about reasons? As long as each of our nine agents
responsible self must allow us to think of other selves from our own perspective rather than theirs. This does not mean that another’s perspective cannot play a role in the way we think of him—it is perfectly possible for me to acknowledge that you have a perspective, and that things seem a certain way to you from it, from my own quite different perspective—but it does mean that we must be able to think of the other without actually occupying his perspective. Because no one has access to anyone else’s subjectivity, any account that identified each responsible self with his own subjectivity would not be truly interpersonal. By contrast, we will have little difficulty understanding how responsibility can be interpersonal if we identify each responsible self with the causes of his subjectivity; for to gain access to these causes, we need not actually occupy the other’s point of view, but need only hold reasonable beliefs about its existence and contents.

A further (and closely related) problem with the first-person approach is that it does not locate responsible selves in the world. If we adopted that approach, and viewed each responsible self as constituted exclusively by his subjectivity, then our conception of such selves would be entirely nonempirical. Because each person’s consciousness is his alone, its particulars are not knowable through the methods of any branch of psychology, folk or experimental, or of any of the physical sciences. By contrast, an account that identifies responsible selves with causally effective psychological structures is, at least in principle, responsive to evidence about the nature of the causally operative states. Because it allows us to introduce such evidence, such an account is suitably interpersonal. For this reason, the third-person approach seems more able than its first-person rival to integrate the moral notion of responsibility into a broadly naturalistic picture of the world.

Even by themselves, these considerations tell heavily against construing responsible selves simply as conscious centers of will. However, the case against that view becomes even more compelling when we turn from the reasons not to privilege the first-person perspective to a careful

III

I have proposed that we identify each agent with the collection of physical and psychological states whose elements interact to sustain his characteristic patterns of conscious and rational activity. If this proposal can be defended, it will provide PEC’s condition (2b) with the theoretical backing it requires. However, I have not yet said anything about why we should adopt the proposal, and so my next task is to provide the needed defense. Put most simply, my strategy will be to argue that each more minimal conception is subject to various difficulties which are best resolved by augmenting it with a causal component.

Let us begin with the version of minimalism that focuses exclusively on the responsible agent’s subjectivity. Because each person is responsible only for his own actions, and because each person’s decisions reflect his own view of his situation, it is quite natural to base our conception of the responsible self on the picture that each of us has of himself when he deliberates. Moreover, although Hume was right to observe that we never encounter the bearer of our thoughts, but only their contents, when we introspect, we nevertheless cannot help thinking of ourselves (and so by analogy others) as standing apart from our thoughts and choosing among the options they present. Thus, if we take as our point of departure the responsible agent’s own perspective on the world, we will naturally think of such agents as simple conscious centers of will.

But should we make the first-person perspective as our point of departure? Doubts arise when we remind ourselves that the concept of responsibility is itself interpersonal. As we saw in chapter 3, we judge that others are responsible at least as often as we render such judgments about ourselves. Because any adequate account of the responsible self must make sense of both sorts of judgments, our concept of the

examination of what that perspective actually involves; for far from compelling us to think of ourselves as constituted exclusively by our own subjectivity, the first-person perspective actually gives us reason to move toward, if not all the way to, a view of ourselves that is much like the one I am defending. To bring out the aspects of the first-person perspective that press us to think of ourselves as constituted partly by causally effective psychological structures, I will begin by briefly revisiting a short stretch of my own recent mental history.

Earlier this morning, when I sat down to write, I did so with the aim of illustrating my claim that our experience as practical agents points to aspects of us that are hidden from consciousness, but I didn’t yet have any particular examples in mind. But why, if I viewed myself simply as a conscious center of will, should I have had any expectation of finding suitable examples of which I was not yet aware? Why was I so confident that I could come up with what I needed? The answer, as I now reconstruct it, is partly that I was counting on tapping into the residue of thought that had been deposited in the months when my mind had circled around different ways of developing the book’s later argument—a residue of whose exact contents I was at the moment unaware, but of whose existence I was nevertheless certain—and partly that experience has taught me to trust my mind’s ability to generate new thoughts more or less on cue. However, to whatever extent my intention rested on the assumption that my previous animadversions had left traces that were not currently before my consciousness, I must, in forming it, have regarded myself as having dimensions that neither were illuminated by the searchlight of my consciousness nor were any part of the illuminative apparatus. Analogously, to whatever extent my intention rested on the assumption that my mind was capable of offering up new thoughts, I must, in forming it, have regarded myself as possessing generative capacities that are unrelated to my will. Thus, taken together, the two explanations of my confidence suggest that I must have viewed myself as far more than a conscious center of will.

This example is of course heavily autobiographical. Because of this, and also because the activity of writing philosophy is highly idiosyncratic, the example may appear to provide my position with only limited support. However, in fact, the example’s details are there only to make it vivid. To generalize the point in a way that shows why practical agents can never view themselves simply as conscious centers of will, we need only remind ourselves of just how restricted the contents of anyone’s consciousness are.

For even if we agree that those contents can include both what an agent is focusing on and what he is passively aware of—even if the driver who was introduced in Chapter 1 can simultaneously be conscious not only of the approaching exit that he is determined not to miss, but also of the car that is approaching on the left, the fact that he is slightly exceeding the speed limit, and his passenger’s unfolding anecdote—these items still represent only the tiniest fraction of the information to which he must take himself to have access in order to carry out his task. If that driver is like most of us, then he is not thinking about any (let alone all) of the turns that he will have to make, or the mechanical operations that he will have to execute, in order to deliver his passenger safely to the airport, and still less is he rehearsing what he will do if he has a flat, if he encounters a traffic jam, if his passenger has a seizure, or if he finds himself in any of the other innumerable unexpected circumstances that could conceivably arise. In each case, he simply trusts himself to think of what is needed when the time comes. No agent could function at all if he did not have confidence that his mind will, just of its own accord, dip into his memory bank to deliver up just the information he needs at just the time he needs it. However, if this reliance on what is available to us despite our not being conscious of it is built right into our practical deliberation (and, I might add, into our theoretical ratiocination) at every step of the way, then the conception of ourselves that informs our deliberation and ratiocination must itself have a substantial nonconscious dimension. This means that thinking of oneself exclusively as a conscious center of will is just about as coherent as thinking of oneself as a front with no back.

By unpacking the assumptions that make it possible to project our practical deliberations into the future, we arrive at a picture of the deliberating self that construes it as (a) containing a substantial nonconscious component, and (b) playing a causal role in generating the thoughts that rise to consciousness at different moments, and (c) generally (though not invariably) functioning in ways that sustain our reason-related activities. In each respect, that picture is at worst consistent with, and at best highly suggestive of, the one I am defending. This means that even if we accept the dubious assumption that the best way to understand the nature of the responsible self is to rely on the picture
of himself that each agent has when he deliberates, we will have little reason to accept, and much reason to reject, the view that responsible selves are simply conscious centers of will.

IV

The minimalist view of the self against which I have just argued has a long history. It is, for example, a recognizable descendant of Descartes’ claim that he is simply a thinking thing. By contrast, my other minimalist target, which takes responsible agents to be constituted exclusively by their ability to form judgments about reasons, is less familiar and of more recent vintage. Thus, before I proceed to my argument, a bit of stage-setting is in order.

As I have said, I think a view of this sort is implicit in an influential approach to responsibility that has recently been developed by Thomas Scanlon and Angela Smith. Scanlon and Smith both maintain that agents are responsible for precisely those attitudes and actions that reflect their judgments about what they have reason to believe or do. They both maintain, as well, that many such “judgment-sensitive” attitudes and actions are not consciously chosen. In a passage that combines both points, Scanlon has written that

The idea of judgment sensitivity helps to isolate the sense in which attitudes can be things we are “responsible for” even when, unlike most voluntary acts, they are not the result of choice or decision. Not only many perceptual beliefs, but many other attitudes as well arise in us unbidden, without conscious choice or decision. Nonetheless, as continuing states, these attitudes are “up to us”—that is, they depend on our judgment as to whether appropriate reasons are present. Because of this dependence on judgment, these are things we can properly be “held responsible” for in several senses of that phrase: they can be properly attributed to us, and we can properly be asked to defend them—to justify the judgment they reflect.7

And Smith expresses the same combination of ideas when she writes:

9. In his influential essay “Two Faces of Responsibility” (in his Agency and Accountability [Oxford: Oxford University Press, 2004]), Gary Watson distinguishes between responsibility as attributability and responsibility as accountability. Because an agent is responsible in the attributability sense when his act flows from, or expresses, his “evaluative commitments,” the first of these notions is at least a close relative of the Scanlon–Smith view. Others who use the term “attributivism” to refer to that view include Neil Levy and Angela Smith herself.
11. In particular, she comes close to endorsing it in “Character, Blameworthiness, and Blame: Comments on George Sher’s In Praise of Blame,” Philosophical Studies 137 (January 2008), 31–39. One of Smith’s targets in that essay is the claim, advanced in my earlier book In Praise of Blame (Oxford: Oxford University Press, 2006), that what connects a wrongdoer to his bad act in a way that makes it reasonable to blame him for it is the causal connection that obtains between his desires, beliefs, and dispositions and the act’s bad-making features. To this claim, Smith objects.

7. Scanlon, What We Owe to Each Other, 21–22.
Like the claim that responsible agents are simply conscious centers of will, the claim that they are constituted by their ability to make judgments about reasons is vulnerable to certain objections that are best met by augmenting this view with a causal component. One such objection emerges when we try to square the attributionist's claim that an agent's responsibility is restricted to those features of his acts that reflect his judgments about reasons with the fact that agents often seem responsible for acts whose wrong-making features have not registered with them at all. So, for example, although Alessandra's extended stay in the school is backed by her judgment that she has good reason to hash things out with the authorities, the feature of her act that makes it wrong—the fact that she is leaving Sheba unattended in a hot car—plays no role at all in shaping this judgment. Because Alessandra has entirely forgotten about Sheba, the dog's plight does not enter her deliberation even as a countervailing consideration whose weight she takes into account but judges to be insufficient to sway her decision. However, if the wrong-making feature of what Alessandra does has no input at all into her all-things-considered judgment that she has reason to do it, then it is hard to see how that judgment can possibly connect her to the act's wrongness in a way that renders her responsible for it.

To deal with cases of this sort, those who restrict an agent's responsibility to those features of his acts that reflect his judgments about reasons sometimes point out that when an agent fails to notice a certain feature of what he is doing, his cognitive failure may itself reflect a judgment that the relevant feature is not a significant source of reasons. Thus, to quote Angela Smith again:

if one judges some thing or person to be important or significant in some way, this should (rationally) have an influence on one's tendency to notice factors which pertain to the existence, welfare, or flourishing of that thing or person. If this is so, then the fact that a person fails to take note of such factors in certain circumstances is at least some indication that she does not accept this evaluative judgment.  

Given the hedged nature of Smith's claim, I see no reason to disagree with it. However, precisely because the claim is so hedged, it does not rule out the possibility that there are also failures to notice morally relevant features of acts that are not themselves judgment-based. In particular, in Alessandra's case (and, I would argue, in many others), the urgency of the dispute, and its high emotional volume, seem by themselves to be quite sufficient to explain her failure to notice her act's wrong-making feature. Because the adequacy of this explanation would not block the conclusion that Alessandra is responsible for leaving Sheba in the car, the essential difficulty remains unresolved. To resolve it, we must locate the significance of Alessandra's failure to remember Sheba not in what it reveals about her judgments about reasons, but rather in its being caused by the same psychophysical structure that sustains her ability to make such judgments.

We can also bring out the need to augment attributionism with some sort of causal component in another way. In its most straightforward form, attributionism takes agents to be responsible for all and only those features of their attitudes and actions that reflect their judgments about reasons. However, when it is understood in this way, attributionism is too straightforward; for we can easily imagine agents whose actions, beliefs, and feelings are indeed grounded in their judgments about what they have reason to do, believe, and feel, yet whom no one would view as responsible for anything they do. To envision such a person, we need only imagine someone whose judgments about reasons are sufficiently capricious or

Why should I blame the fact that some random collection of psychological states have causally interacted in my mental history in such a way as to produce a bad action? What do I have to do with any of that? (33)

Smith also writes that "we seem to have lost the agent in all of this talk about the desires, beliefs, and dispositions that operate within him to produce his actions" (33) and that

[i]n my view, the most plausible candidate for what it is that links an agent to his attitudes and dispositions—as well as to his intentional actions—is his evaluative judgment…. It is these judgments—judgments about what is good, worthwhile, or important—that in my view provide the crucial "link" between the agent and his attitudes and actions (34).

Taken together, these passages strongly suggest that Smith takes judgments about reasons to be relevant not only to the conditions under which we can attribute responsibility but also to the nature of the agents to whom we can attribute it.

12. Smith, "Responsibility for Attitudes," 244.
unintelligible. Suppose, for example, that someone takes the fact that it is raining to be a reason to shave the left side of his head, the fact that the floor is not swept to be a reason to expect a financial windfall, and the fact that his cat's nose is running to be grounds for intense envy; and suppose, further, that these strange judgments are not backed by any further beliefs that are not equally implausible. Although this agent's actions, beliefs, and feelings will all be grounded in judgments about reasons, that will hardly qualify him as responsible; and the reason is clearly that his judgments about what he has reason to do, believe, and feel will bear no relation to what he in fact has reason to do, believe, and feel.

Because what makes this case problematic is the agent's inability to recognize his actual reasons, the obvious way for the attributionist to block the difficulty is to include that ability in the features that someone must have to be a responsible agent. Instead of taking responsible agents to be constituted by their ability to form judgments about what they have reason to do, believe, and feel, he must take them to be constituted by the more demanding ability to form accurate judgments about what they have reason to do, believe, and feel. If the need for this addition is not immediately apparent, it is probably because most of the judgment-sensitive attitudes and actions that we encounter are grounded in judgments about reasons that we can at least understand even if we do not share them. However, even if we rarely deal with agents whose judgments about reasons float entirely free of reality, the fact that such individuals can and do exist, and that we are definitely not inclined to regard them as responsible, makes it hard to see how any attributionist could resist such some addition.

But if the attributionist does not resist the addition, then he may also have to introduce a causal element into his account. The introduction of such an element may be necessary because it is hard to see how an agent's actual and counterfactual judgments about his reasons could systematically come out to be true without being causally dependent on whatever makes them true. If there were no causal relation between the agent's judgments about his reasons and their truth-makers, then the continuing actual and counterfactual accuracy of those judgments would simply be a miracle. Because a capacity to reach accurate judgments about one's

reasons is only intelligible if we take it to have a causal basis, anyone who takes that capacity to be constitutive of responsible selves must be under heavy pressure to take its causal basis to be constitutive of them too.

And, because of this, the need to incorporate the capacity to recognize one's actual reasons will move attributionism much closer to my own account. The reason for the convergence is not just that the new attributionist claim that responsible agents must be capable of forming accurate judgments about their reasons will coincide with my own earlier own claim, advanced in chapter 7, that even someone whose failure to recognize that he is acting wrongly or foolishly is due to his constitutive psychology will not be responsible for acting that way unless he has the general capacity to draw the sorts of conclusions from his evidence that he did not draw on this occasion. It is, more deeply, that my broader account offers just the resources that are needed to capture the causal basis of any such capacity. Because I have proposed that we identify each responsible agent with whatever psychological or physical structure gives rise to his judgments about what he has reason to believe or do, I can build the accuracy requirement into my account by simply adding that any relevant structures must have causal properties that render the agents' judgments about reasons accurate within an appropriate range of actual and counterfactual situations. Indeed, because I originally identified the relevant causal structures by working backward from instances in which responsible agents do recognize and respond to their actual reasons, this requirement has been built into my account from the start. Thus, by following up on the need to tether the responsible agent's judgments about reasons to reality, we encounter yet a further reason to move in the direction of an account like mine.

V

To answer the linked questions of what makes someone a responsible agent and what makes someone the particular responsible agent he is, I have argued that we must look beyond the agent's consciousness and reason-responsiveness to the causal structures that sustain them. The resulting account is far more inclusive than the minimalist views which hold that what makes someone a responsible agent is simply
his consciousness and/or reason-responsiveness and that what makes someone the particular responsible agent he is, is simply the contents of his consciousness and/or his judgments about reasons; but it is far less inclusive than the maximalist view that treats all facts about a responsible agent as equally constitutive of him. Mine is, however, far from the only account to occupy this intermediate territory, and so I must say something about its relation to its conceptual neighbors.

We can divide the most important of these into two broad groups, both of which also take responsible agents to be constituted by some but not all elements of their psychology, but which disagree about which elements are in and which are out. Put most simply, the difference is that one group draws the distinction synchronically while the other draws it diachronically. Views of the first sort, inspired by the influential work of Harry Frankfurt, take each responsible agent to be constituted by just those aspects of his character that he in some sense accepts. Philosophers who take this approach explicate the relevant form of acceptance in terms of factors such as wanting to have and be moved by one’s lower-order desires, holding values that endorse or at least do not condemn those desires, identifying with certain attitudes while dissociating oneself from others, and being unable to avoid these forms of identification or dissociation. By contrast, views of the second sort seek to exclude attitudes with objectionable causes such as manipulation, brainwashing, and certain forms of conditioning by restricting a responsible agent’s constitutive features to those aspects of his character that have the right kind of causal history.

To say that a responsible agent is constituted only by attitudes he accepts is not to say that he is constituted by all such attitudes, so we can easily combine versions of the two approaches. To do so, we need only say that being accepted by the agent is one necessary condition for inclusion while having the right kind of history is another. Taking his cue from this, someone might be tempted to go further and combine versions of either or both accounts with my own. However, it is important to realize that, for simple logical reasons, this last option is not open.

To see why this is so, let us first consider the combination of my own account, which asserts that each responsible agent is constituted by the full set of psychological states whose interaction sustains his rationality-related activities, with the view that responsible agents are constituted only by those psychological states that they in some sense accept. If someone were to advance this combined account, then he would have to acknowledge that there could be a person for whom one and the same psychological state was both an element of the psychological structure that causally sustained his rationality-related activities, and thus was constitutive of him, but was also an aspect of his personality that he rejected, and thus was not constitutive of him. Because this conjunction is a straightforward contradiction, no combination of views that implies its possibility can be accepted. Moreover, because the set of psychological states that sustains an agent’s rationality-related activities might well operate very differently, or might not operate at all, if we subtracted some of its elements, we cannot avoid the difficulty by stipulating that an agent’s constitutive features consist only of whichever elements of the sustaining set he does not reject. Because of this, and because a similar contradiction would emerge if we tried to combine my own account with the view that a responsible agent is constituted only by that subset of his psychological states with the right kinds of histories, my account is clearly incompatible with each of its more famous alternatives.

Although this implication obviously raises the stakes, the escalation is not one by which I am particularly bothered; for neither alternative appears to me to be especially strongly motivated. Where the ‘acceptance’ view is concerned, the difficulty is partly that describing an agent’s aversion to a hard-to-resist and destructive element of his personality in the language of alienation and otherness seems distorting and tendentious—isn’t it at least as accurate to describe him as appalled that he is that way?—and partly that even the strongest feeling that a certain element of one’s personality is alien or external just doesn’t seem to be a very good basis.

---

14. Harry Frankfurt introduces this idea in his seminal essay “Freedom of the Will and the Concept of a Person,” and explores it in a number of the other papers in his collection The Importance of What We Care About (Cambridge: Cambridge University Press, 1988).

upon which to conclude that it really is. Where the right-kind-of-history view is concerned, the difficulty is that the simplest explanation of why those who were brainwashed, indoctrinated, and the like do not seem responsible is that they have been rendered incapable of responding to certain sorts of reasons—a fact that has nothing to do with their histories and everything to do with their current mental makeup. These observations are of course the merest gestures at arguments, and I proffer them not as reasons to reject the views with which mine is incompatible, but only as explanations of why I am not bothered by the incompatibility. My reason for rejecting these views is the one I have already given—that by taking as our point of departure the consciousness and reason-responsiveness that responsibility undeniably presupposes, we are led to adopt the sort of causal account that rules them out.


At the end of Chapter 6, I remarked that given the searchlight view’s unacceptable implications, the real question was not whether that view had to go, but only whether the concept of responsibility could survive its excision. Now, with my developed alternative, PEC, before us, I want to end by returning to this question. Because PEC applies exclusively to acts that are wrong or foolish, one thing we must ask is whether it can be integrated into a larger account that also applies to right and prudent acts. More fundamentally, because PEC detaches an agent’s responsibility from his awareness of what he is doing, there are questions about its consistency with both responsibility’s voluntariness condition and the widely held view that responsibility requires control. In this final chapter, I will take up each of these challenges. Although they all raise interesting issues about responsibility, I will argue that none of them show the concept to be incoherent.

I

Let us begin with the distinction between acts that are wrong or foolish, on the one hand, and acts that are right and prudent on the other. As we have seen, it is natural to suppose that there is a single concept of responsibility