GARY WATSON: STRAWSONIAN

Michael Smith

In the subtitle of his "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme" (Watson 1987), we learn that Gary Watson self-conceives as someone whose views about moral responsibility have been heavily influenced by P. F. Strawson's "Freedom and Resentment" (Strawson 1962). Given that in his early work Watson defends claims that, if true, would show that Strawson's view about moral responsibility is vulnerable to collapse from within, I must confess to being surprised when I first read that subtitle, but it also piqued my interest. I was keen to learn why Watson should think of himself as a Strawsonian.

You can therefore imagine my surprise when I finally read Watson's paper and came across a passage towards the end that crystalized for me what I had always found so implausible about Strawson's view, and further discovered that Watson thinks that that feature of Strawson's view is implausible too, though apparently not so implausible as to make him reject Strawsonian views altogether. My aim in what follows is thus to talk through how Strawson's and Watson's views of moral responsibility relate to each other. In the course of doing so, a more positive view will emerge. Whether that more positive view is itself Strawsonian is a matter that I will leave to others to decide. In some respects it is, in others it isn't.

1. Strawson's view of moral responsibility

Familiarly enough, Strawson begins "Freedom and Resentment" by contrasting two views about the problem that causal determinism poses for our practice of holding each

other responsible. He calls those who hold the first of these views 'pessimists' and those who hold the second 'optimists'.

According to the pessimists, our practice of holding each other responsible presupposes that we are responsible for doing something only if we have the capacity to do otherwise, a capacity that causal determinism would show that we lack. This is because if causal determinism is true, the thing that we do is the only thing that we can do. According to the optimists, by contrast, our practice of holding each other responsible requires no such counter-causal capacity. It requires only that, by holding each other responsible, we thereby regulate each other's conduct. The truth of causal determinism is irrelevant, according to this second view, because whether or not our holding each other responsible has such regulatory effects is independent of the truth of causal determinism.

While Strawson agrees with the optimists that the truth of causal determinism is irrelevant to our practice of holding each other responsible, he rejects the optimists' view for the same reason that he rejects the pessimists'. Both assume that something external to our practice of holding each other responsible explains why it is correct for us to do so. Pessimists assume that it is correct for us to hold someone responsible only if they have the capacity to do otherwise. Optimists assume that it is correct for us to hold someone responsible only if we are suitably responsive to being held responsible for our conduct. Strawson's view, by contrast, is that nothing external to our practice of holding each other responsible explains why it is correct to do so. Instead, our holding each other responsible is appropriate just in case various conditions that are internal to the practice of holding each other responsible are met. What are these internal conditions? According to Strawson, these emerge when we carefully describe the practice.

Restricting our attention to responsibility for causing harm, the relevant features of the practice turn out to be the following:

> (i) When we hold each other responsible, we typically have some reactive attitude or other towards those we hold responsible: guilt or shame, in the case of our holding ourselves responsible for harming someone; resentment or anger, in the case of our holding someone responsible who harms us; and indignation or anger, in the case of our holding someone responsible who harms a third party.
>
> (ii) These reactive attitudes are all natural human responses to the quality of will that people display when they cause harm, responses that in turn get expressed in the demand that we display a suitable degree of good-will towards each other.
>
> (iii) Though natural human responses, the reactive attitudes are rationally grounded and so can be defeated in certain circumstances. Imagine that there was a justification for the harm done, and the person who caused the harm knew of that justification—perhaps it was done to prevent a much worse harm. Or imagine that the person who caused the harm had an excuse—perhaps the harm was done by a child who doesn't understand what it is to harm someone, or by someone who does understand that, but couldn't tell that they were harming someone because they had been misled, or by someone who was temporarily or permanently insane or in some other way mentally impaired. If we are rational, then once we become apprised of such a justification or excuse our reactive attitudes will disappear or be moderated.

The upshot is that, when we carefully describe the practice of holding each other responsible, we discover two things, one negative and one positive.

Negatively, we discover that when we hold someone responsible, we have certain reactive attitudes towards them, and that though these attitudes can be defeated in certain circumstances, living in a world in which causal determinism is true is not one of those circumstances. It provides us with neither a justification for causing harm nor an excuse. In this respect, Strawson's view is like the optimists'. However, unlike the optimists, Strawson thinks that the reactive attitudes are sensitive to issues of control—in this respect his view is more like the pessimists'—but the relevant idea of control is the everyday idea, not the pessimists' counter-causal idea. Do those who cause harm have the capacity to understand what they're doing? If so, do they have ordinary capacities for reasonable agency, capacities that are compromised in some way when people are misled, or insane, or in some other way mentally impaired? If the answer to these questions is 'yes', then those to whom we have the reactive attitudes have no excuse. If they also lack a justification for doing what they did, then it is perfectly appropriate for us to have the reactive attitudes that we naturally have towards them.

More positively, what we discover when we carefully describe our practice of holding each other responsible is that that practice itself tells us when it is correct to hold people responsible, and that this in turn provides us with all we need in order to explain what it is for them to be responsible. Someone is responsible just in case it is correct to hold them responsible, and it is correct to hold someone responsible just in case we are naturally disposed to have certain reactive attitudes towards them for doing what they did when they lack either a justification or an excuse. Since these conditions are often met, Strawson thinks of this as a vindication of our ordinary practice of holding each other responsible.

Armed with this more positive account of when people are responsible, we can be more precise about where Strawson thinks the optimists and pessimists go wrong. The pessimists think, mistakenly, that we can understand what it is to be responsible in terms of the idea of control, but that this has to be understood counter-causally, and so independently of what it is to hold someone responsible. According to Strawson, by contrast, the relevant idea of control can only be understood in terms of the conditions under which the reactive attitudes are defeated, that is to say, in terms of the idea that those to whom we are naturally disposed to have certain reactive attitudes lack a justification or an excuse for doing what they did. The optimists thus go too far in thinking that the idea of control is irrelevant to our understanding of what it is for someone to be responsible. Even if we could regulate people's conduct by holding them responsible for doing what they did, he thinks that that wouldn't show them to be responsible, as we might still not be disposed to have certain reactive attitudes towards them for doing what they did when they lack either a justification or an excuse.

2. The bearing of Watson's early work on Strawson's view

How convincing is Strawson's view? As I understand it, Watson's early paper "Skepticism about Weakness of Will" poses a significant challenge to that view (Watson 1977). Watson's aim in that paper is to describe in some detail our ordinary practice of holding people responsible, and in particular the ways in which we ordinarily suppose that the capacities required for reasonable agency can be impaired. In so doing he argues that we ordinarily make distinctions where there are no differences. If he is right, then our ordinary practice of holding people responsible is defective for reasons having nothing to do with causal determinism.

Watson focuses on cases of our having, but failing to exercise the capacity to form desires in the light of our beliefs about what we have reason to do: that is, cases of weakness of will, as that is standardly understood. Imagine two people who both know that they have a decisive reason to act so as to avoid harm, but don't desire to act in that way, where one's lack of desire is the result of their having but failing to exercise a capacity to desire in accordance with their knowledge, and the other's is the result of their incapacity to desire accordingly. Though they both lack self-control, the second has no self-control to exercise, whereas the first does, but simply fails to exercise it. Assuming that both act on the desires they in fact have and cause harm, we would ordinarily suppose that the first meets a condition necessary for being held responsible, whereas the second does not.

Note that Watson's concern readily generalizes to the capacity for belief formation. Imagine two people who both falsely believe that they have decisive reason to act in a way that causes harm, where one has that false belief as a result of their having but failing to exercise the capacity to believe truly that they have no such decisive reason, and the other has that false belief as a result of their lacking the capacity to believe the truth in the first place. The second person is cut off from reality, whereas the first isn't. He has the wherewithal to access reality, but fails to make use of it. If both of these agents act on their false beliefs, then we would ordinarily suppose that the first meets a condition necessary for being held responsible, whereas the second does not.

As Watson sees things, the problem with our making these ordinary distinctions is that we cannot come up with a plausible folk psychological explanation of why the person who has a capacity but fails to exercise it in circumstances like those described

fails to exercise that capacity. Consider those cases in which an agent believes that he has a decisive reason to act so as to avoid harm, and has a capacity to desire to act in that way, but doesn't exercise that capacity and acquire the desire to act accordingly. In typical cases like this the agent's failure to exercise their capacity is explicable folk psychologically. Perhaps they changed their mind about what they have a reason to do, or perhaps they were incapable of desiring to act accordingly. But neither of these explanations is supposed to be available in cases like that described, as these are meant to be cases of weakness of will. But in that case what does explain their failure? The problem, as Watson sees things, is that we need a folk psychological explanation of the agent's failure to exercise their capacity in such cases, but we lack a folk psychological explanation that leaves the facts of the case intact. Our ordinary practice of holding people responsible is in this way shown to invest normative significance in a distinction where we can find no difference.

A variation on the same problem arises in those cases in which an agent falsely believes he has a decisive reason to cause harm despite the fact that evidence to the contrary is available to him. What explains the agent's failure to avail himself of that evidence? One possibility is that he lacks the capacity to avail himself of that evidence, or that he has that capacity but lacks the capacity to change his belief in the light of the evidence he avails himself of. But since these are meant to be cases in which we hold agents responsible for their false beliefs, neither of these explanations is supposed to be available. So what does explain the agent's failure to change his mind? Once again, the problem Watson identifies is that we need a folk psychological explanation of the agent's failure to exercise their capacity, but we lack an explanation that leaves the facts of the

case intact. Our ordinary practice of holding people responsible is once again shown to invest normative significance in a distinction where there is no difference.

If Watson is right, then this is very bad news indeed for Strawson. Strawson assumes that if our ordinary practice of holding each other responsible is immune to attack externally on the grounds of causal determinism, or on the grounds that our holding each other responsible doesn't have certain regulatory effects, then that amounts to a vindication of the ordinary practice. But if Watson is right then this assumption is unwarranted. Our ordinary practice of holding each other responsible may still lack a vindication because it cannot meet standards of internal coherence to which it must also be held. If the distinction between an agent's having a capacity but failing to exercise it and lacking a capacity altogether is a distinction without a difference, then the ordinary practice is vulnerable to attack on precisely these grounds, as it turns out that we regularly hold certain people responsible and let others off the hook without having a good reason for our differential treatment of them.

Watson's own view is that, despite this flaw in our ordinary practice of holding each other responsible, we can reconstruct our practice around a related distinction that tracks the normative judgements we ordinarily make. The related distinction requires us to divide the agents we are disposed to hold responsible for causing harm into two classes. There are those we are disposed to hold responsible who have normal powers of self-control, and there are those we are disposed to hold responsible who lack such powers. Even though when they each cause harm, neither has the capacity to act otherwise, we should think that those in the former class have no excuse for doing what they do, whereas those in the latter class do have an excuse. This requires a

reconstruction of our ordinary practice because the feature that those in former class have which prevents them from having an excuse isn't that they have but failed to exercise the capacity to exercise self-control—no one has that—but rather that they lack normal powers of self-control.

This view should sound familiar, as in its essentials it is the same as the explicitly Strawsonian view that Jay Wallace defends in his *Responsibility and the Moral Sentiments* (Wallace 1994). Wallace defends his view on normative grounds. He tells us that our ordinary standards of fairness don't rule out our holding those in the former class responsible, whereas they do rule out our holding those in the latter class responsible. This suggests a Strawsonian direction of explanation. People in the former class are responsible because it isn't unfair to hold them responsible, rather than vice versa. Facts about our practice of holding people responsible are in this way still prior to the facts about who is and isn't responsible. To the extent that Watson self-conceives as a Strawsonian about what it is for someone to be responsible, I take it that he does so because he holds a somewhat similar view.

3. Why Watson is wrong about unexercised capacities

What should we think about Watson's alternative account of what it is for someone to be responsible? To my mind, two crucial points need to be made about it (Smith 2003, for different but related responses see Mele 2012 and McGeer and Pettit 2015).

The first is that Watson's demand that we give a further folk psychological explanation of an agent's failure to exercise a capacity he possesses is unreasonable. When someone does something because they fail to exercise a capacity they possess, all there may be to say folk psychologically is that though they had that capacity, and though

its exercise was called for, and though in certain cases like those of weakness of will they may have known that this was so, they simply failed to exercise the capacity that they had. Since this will show them to be irrational, there is a sense in which such explanations fall short of those we aspire to give when we usually give folk psychological explanations. Many folk psychological explanations are attempts to show that what the agent did was rational in their circumstances. But when an agent has the capacity to desire to do what he believes he has reason to do and fails to exercise it when its exercise is called for, or when he has the wherewithal to avail himself of evidence that would lead him to change his beliefs but he doesn't avail himself of that evidence, there is no way to show that the agent responded rationally to his circumstances. In such cases, and indeed in all similar cases of irrationality, a folk psychological explanation can amount to no more or less than a description of irrational state of mind that led the agent to do what he did. To insist on something more or different, as Watson does, is itself to distort the facts of the case.

The second point to make in response to Watson's alternative account of what it is for someone to be responsible is that he is wrong the distinction between those who have a capacity but fail to exercise it, and those who lack a capacity altogether, is a distinction without a difference. We can spell out exactly what difference we have in mind when we make this distinction. The claim that someone has but fails to exercise a capacity entails the modal claim that they could have done something that they failed to do. It might be thought that this is what's shown to be impossible if causal determinism is true, but the modal claim need not, and in my view should not, be understood in these counter-causal

terms. Instead we should spell it out in the same way that we spell out other modal claims.

Roughly speaking, the agent who has but fails to exercise a capacity is surrounded by nearby possible worlds in which he succeeds, whereas the one who lacks the capacity is not. A little more precisely, if we imagine two people, one of whom has a capacity but fails to exercise it, and the other of whom fails to exercise the capacity because he lacks it, then if we consider the nature of the possible worlds in which they each fail, and the nature of the nearest possible worlds in which they each succeed, and then compare how similar the failure and success possible worlds are to each other along a relevant dimension of similarity, what we find is that the degree of similarity in the case of those who have but fail to exercise the capacity is greater than it is in the case of those who lack the capacity. This is what we mean when we say that an agent has a capacity, but fails to exercise it.

What is the relevant dimension of similarity? Since the agent who has but fails to exercise the capacity possesses but fails to exercise powers of self-control, these are the nearest possible worlds in which the agent exercises the powers of self-control that he in fact has. The nature of these possible worlds is fixed by the imaginative and attentional resources the agent employs when he doesn't give into temptation, in the case of weakness, or when he does avail himself of evidence, in the case of revising his beliefs. Since these imaginative and attentional resources will presumably vary from agent to agent, and within an agent from time to time, it would be a difficult task indeed to describe these nearby possible worlds with any generality. The crucial point, however, is one that we can make without providing such general descriptions, namely, that there are

no such facts to constrain the nearby possible worlds in which agents who lack powers of self-control exercise their powers. The nearest possible worlds in which they exercise their powers of self-control are therefore both less similar to the failure worlds and less similar to each other than are the possible worlds in which the agent who has such powers exercises his.

Note that this two-part response to Watson acknowledges what's right about his positive view. Watson is right that the fact that some agents normally exercise self-control and others don't is normatively relevant to our differential treatment of them. But he is right not because this allows us to provide a rational reconstruction of our justification for treating them differently, but rather because it allows us to explain the ordinary justification for treating them differently. Those agents who have but fail to exercise a capacity lack an excuse for doing what they did and so are at fault, and those who lack the capacity altogether do have an excuse and so aren't at fault. The two-part response also acknowledges that there is something right about Strawson's view. Strawson insists that the relevant idea of control that's needed for an agent to be held responsible isn't counter-causal, but is rather grounded in facts that are internal to our practice of holding each other responsible. This turns out to be true too, for internal to that practice is the fact that certain agents usually employ strategies of self-control whereas other agents don't, and whether or not we hold agents responsible is sensitive to this difference between them.

However the news for Strawson isn't all good, as the two-part response also suggests that we no longer need to explain what it is for someone to be responsible in the way he does, that is, in terms of when it is correct to hold them responsible. We can

instead suppose that it is correct to hold someone responsible when they are responsible. Imagine someone who causes harm. They are responsible for doing so when they had a decisive reason not to do so—this is the echo of Strawson's no justification condition—and when the explanation of their causing harm is that they had, but failed to exercise the capacity to control themselves—this is the echo of Strawson's no excuse condition: as we put it above, they are at fault. Armed with this account of when agents are responsible, we can then suppose that it is correct to hold them responsible when this condition is met, incorrect otherwise.

4. Additional problems for Strawson and Watson

So far I have been concerned to spell out Strawson's view of moral responsibility and to explain how someone who adopts that view can respond to the challenge posed to it by Watson's early work. I have also suggested that, equipped with that response, we can give an independent account of when someone is responsible for acting in a certain way, an account in terms of which we can explain when it is correct for us to hold them responsible for doing so. There are, however, further problems looming for Strawson, problems that Watson brings out towards the end of "Responsibility and the Limits of Evil."

Let's return to Strawson's positive view. To repeat, he thinks that people are responsible just in case it is correct to hold them responsible, and he further thinks that holding someone responsible is a matter of our having certain reactive attitudes towards them, attitudes that get expressed in the demand that those who cause harm show others a suitable degree of good will. Watson calls a theory of this kind an *expressive theory*, and he self-conceives as a Strawsonian in part because he thinks that some version of the

expressive theory is correct. The distinctive feature of Strawson's own version of the expressive theory is the account he gives of the attitudes we express when we hold someone responsible. As he sees things, we hold someone responsible when we have certain reactive attitudes towards them. Further problems for Strawson's view come to light when we examine more closely what these attitudes are like.

Consider the trio of reactive attitudes guilt, resentment, and indignation. If these all get expressed in a demand that someone who caused harm shows others a suitable degree of good will, then they all presumably consist in part in thinking, rather than believing, that that person has caused harm and so violated the demand without having an excuse for doing so—or, more prosaically, in thinking that they have caused harmed when they had a decisive reason not to do so without an excuse. The difference between these reactive attitudes then lies in the fact that, in the case of guilt, the person who is thought to have caused the harm is oneself; in the case of resentment, it is someone else, but the harm is thought to have been caused to oneself; and in the case of indignation, it is someone else and the harm is thought to have been caused to a third party.

These reactive attitudes are all constituted by thoughts rather than beliefs because we can evidently experience guilt, resentment, and indignation even when we know that no demand has been violated (Rosen 2015, Gease 2016). Imagine the guilt you might feel if you ran over a child, and so caused their death, through bad moral luck: you were driving perfectly safely when the child ran onto the road in front of you. Or imagine the anger and resentment that the child's parents might feel towards you. Such reactive attitudes are, of course, irrational given that no demand has been violated, but they are still possible, notwithstanding their irrationality. The suggestion is that having reactive

attitudes like these is possible because we can still find ourselves *thinking* that a demand has been violated.

Importantly, however, having such thoughts is insufficient for our having the reactive attitudes. Focus on the case of indignation. I could well think that someone has caused a third party harm when they had a decisive reason not to, and that they lacked an excuse for doing so, and yet be positively gleeful about that fact, rather than indignant. Imagine that I hear news that a hitman I hired to kill my nemesis has succeeded, but imagine further than I know full well that it is wrong to kill people, and hence that the hitman had a decisive reason not to do what I paid him a significant sum of money to do. What cases like this suggest is that, even if having the reactive attitudes consists in part in thinking that someone has caused another harm when they had a decisive reason not to and lacked an excuse, it must also consist in part in some further attitude. The difficult task is to say what that further attitude is.

I said earlier that a passage in Watson's "Responsibility and the Limits of Evil" crystallized for me what I found so implausible about Strawson's view when I first read his paper. That passage begins with a quote from Strawson's essay in which he provides his account of the further attitude required for having reactive attitudes. Here is that passage with Watson's elisions.

> Indignation, disapprobation, like resentment, tend to inhibit or at least to limit our goodwill towards the object of these attitudes, tend to promote at least partial and temporary withdrawal of goodwill…(These are not contingent connections.) But these attitudes…are precisely the correlates of the moral demand in the case where the demand is felt to be disregarded. The making of the demand is the

> proneness to such attitudes…The holding of them does not…involve…viewing their object other than as a member of the moral community. The partial withdrawal of goodwill which these attitudes entail, the modification they entail of the general demand that another should if possible be spared suffering, is…the consequence of continuing to view him as a member of the moral community: only as one who has offended against its demands. So the preparedness to acquiesce in the infliction of suffering on the offender which is an essential part of punishment is all of a piece with this whole range of attitudes…(Watson 1987: p.90)

Strawson thus thinks of the reactive attitudes as retributive sentiments. The trio of attitudes therefore consists in part in a desire that the person who caused harm when they had a decisive reason not to, and lacked an excuse, suffers proportionally for doing what they did: in the case of guilt, one desires that one suffers oneself; in the case of resentment, one desires that the person who caused one harm suffers; and in the case of indignation, one desires that the person who caused harm to the third party suffers.

Note that if Strawson were right about this then that would indeed explain why I can be gleeful when I think about the hitman having killed my nemesis, rather than indignant. This would be possible because though I can have the thought that's partially constitutive of indignation, I can lack the desire that the hitman suffers proportionally for doing what he did. However if Strawson were right then this would have other far less plausible consequences as well. The most striking of these is that those like myself who lack retributive desires—in my view retributivism is a barbarically false moral doctrine,

and my feelings towards those who do wrong square well with my moral beliefs—would be incapable of holding people responsible.

It was this implication of Strawson's view that I found so implausible when I first read his paper. Strawson is supposed to be providing an analysis what it is hold someone responsible, an analysis that should be compatible with those who hold one another responsible having a variety of views about what the demands of morality are. But his analysis of what it is to hold someone responsible builds in his own (to my mind) false views about what morality demands us to do to those who violate its demands. Since the claim that someone is responsible for causing harm, in the sense that Strawson assigns to that notion, presupposes the truth of the barbarically false moral doctrine that he embraces, it follows that no one is responsible in the sense he specifies.

Watson's reaction to the passage from Strawson's essay suggests that he more or less agrees with this assessment.

> This passage is troubling. Some have aspired to rid themselves of the readiness to limit goodwill to acquiesce in the suffering of others…out of a certain ideal of human relationships, which they see as poisoned by the retributive sentiments. It is an ideal of human fellowship or love which embodies values that are arguably as historically important to our civilization as the notion of moral responsibility itself. The question here is not whether this aspiration is finally commendable, but whether it is compatible with holding one another morally responsible. The passage implies that it is not.
>
> If holding one another responsible involves making the moral demand, and if making the demand is the proneness to such attitudes, and if such attitudes

> involve retributive sentiments and hence a limitation of goodwill, then skepticism about retribution is skepticism about responsibility, and holding one another responsible is at odds with one historically important ideal of love. (Watson 1987: p.256)

Watson thus thinks that Strawson's own version of the expressive theory faces a dilemma. On the first horn of the dilemma, which is what he describes here, we grant his claim that holding someone responsible for causing harm is a matter of our having one or another of the reactive attitudes towards them, where these are thought of as retributive sentiments, and we conclude that holding people responsible is a far less ubiquitous phenomenon than we had thought. Those who lack retributive sentiments never hold people responsible, and in their view no one is ever responsible in the sense that Strawson specifies to the idea of someone's being responsible in his essay. This is a seriously revisionary view, one that is at odds with the expressive theory's aspiration to vindicate our ordinary practice of holding one another responsible.

On the other horn of the dilemma, the horn that Watson plainly favors, we reject Strawson's claim that holding someone responsible for causing harm is a matter of our having one or another of the reactive attitudes towards them, where these are thought of as retributive sentiments, but we still give a version of an expressive theory. After the passage just quoted in which Watson explains how forswearing the retributive sentiments is one "historically important ideal of love", he puts this horn of the dilemma.

> Many who have this ideal, such as Ghandi or King, do not seem to adopt an objective attitude in Strawson's sense [*MS:* ie they do not limit their goodwill to those who cause harm in the way Strawson suggests is appropriate]...They *stand*

> *up* for themselves and others against their oppressors; they *confront* their oppressors with the fact of their misconduct, *urging* and even *demanding* consideration for themselves and others; but they manage, or come much closer than others to managing, to do such things without vindictiveness or malice.
>
> Hence, Strawson's claims about the interpenetration of responsibility and retributive sentiments must not be confused with the expressive theory itself. As these lives suggest, the retributive sentiments can in principle be stripped away from holding responsible and the demands and appeals in which this consists. What is left are various forms of reaction and appeal to others as moral agents. The boundaries of moral responsibility are the boundaries of intelligible moral address. To regard another as morally responsible is to react to him or her as a moral self. (Watson 1987: pp.256-7)

On this second horn of the dilemma, we hold onto the expressive theory's idea that holding someone responsible is a matter of expressing reactive attitudes, but we give an alternative account of what the non-cognitive component of such attitudes is like.

What Watson actually says in this passage is, of course, less than what's required to fully spell out an alternative version of the expressive theory. He tells us what he thinks the non-cognitive attitudes we have when we hold someone responsible lead us to do, but not what those attitudes are. But it might be thought that we can infer what they are from his account of what they lead us to do. If these non-cognitive attitudes lead us to (say) demand consideration for ourselves and others, then they must at the very least be desires like those George Sher thinks are part of the complex of attitudes we have when we blame someone (Sher 2006). According to this more fully spelled out alternative

version of the expressive theory, holding someone responsible for causing harm would be a matter of thinking that they have caused harm when they had a decisive reason not to do so and lack an excuse, and desiring that they not have caused that harm, and perhaps that they not cause harm more generally.

The trouble with this alternative version of the expressive theory is, however, that it is doubtful that holding someone responsible requires us either to have such desires or to engage in any of the forms of moral address that Watson describes. This is not to deny that when we hold someone responsible we typically are disposed to engage in such forms of moral address; it is simply to insist that we distinguish between this typical connection and the constitutive connection that would be required for the alternative version of the expressive theory Watson describes to be correct. If it isn't obvious that that connection is merely typical rather than constitutive, then I suspect that that's because we rarely consider the full range of cases in which we hold people responsible for causing harm, focusing almost exclusively on cases in which one person is harmed by another. But when we consider other cases, such as those in which an agent causes harm to himself, we see that the idea that holding them responsible for the harm they cause consists in having desires like those described, or in being disposed to engage in forms of moral address such as urging them to show themselves more consideration, is quite implausible.

To fix ideas, imagine that you open a coffee shop, and that the next day a competitor opens a coffee shop next door. Local demand is sufficient to support one coffee shop, but not two. As time goes by, you learn that your competitor is making bad business decisions. He doesn't have sufficient supplies of the various kinds of coffee or

pastries he has on the menu, and he is gruff and unapologetic when his customers complain, so they leave dissatisfied. Over time more and more of his customers come to your shop rather than his, and he eventually goes out of business. Now ask yourself whether you hold anyone responsible for the failure of your competitor's business, and if so, who. I take it that the answer to these questions is obvious. You hold your competitor responsible for the failure of his business. It wasn't just bad luck that he failed, and nor was it anyone else's fault. He caused harm to himself when he had a decisive reason not to do so and, assuming he wasn't self-destructive or suffering from some temporary mental illness, he had no excuse for doing so.

Now ask yourself whether it follows from the fact that you hold your competitor responsible that you must at some point have desired him not to have done what he did, or that you were disposed to (say) urge him to make better business decisions because of the harm he is doing to himself. Once again, I take it that the answer to these questions is obvious. You need have had no such desires and or be so disposed. When you operate a business in a competitive environment, you have decisive reasons to do various things for your competitors. You mustn't sabotage their efforts to make a go of things, and you mustn't give yourself an unfair advantage by breaking local laws that govern the operation of your business. But, contrary to Sher, it seems perfectly plausible and permissible that, in circumstances like these, you might have hoped all along that your competitor makes mistakes exactly like those described, mistakes that would put him out of business, and contrary to Watson, you need have no reason at all to give your competitors a pep talk to help them succeed at your expense.

5. A positive suggestion

This leaves us with something of a problem. If holding people responsible for causing harm isn't, inter alia, a matter of having retributive sentiments towards them, or desiring that they not act in the way they do, then what does it amount to? My own positive suggestion is Watsonian in a certain respect.

It seems to me that Watson is right that to hold someone responsible is to react to them as a moral agent. However my own view of what it is to react to someone as a moral agent draws on the account given earlier in reply to Watson's suggestion that the distinction between someone's having but failing to exercise a capacity, and failing to possess that capacity altogether, is a distinction without a difference. To be a moral agent, it seems to me, is to be someone who has the capacity to do what they have reason to do, where this is a capacity that they may or may not exercise. To learn that an agent is a moral agent is thus to acquire an expectation of them that they will do what they have reason to do to some extent, and, when they fail to do what they have reason to do, it is to think of the possible worlds in which they do what they have reason to do as, in the sense specified earlier, nearby: they had the capacity to exercise self-control, but they failed to exercise it. If we think of trusting someone as a disposition to treat them as if they will do what they have reason to do, then we can put the point in terms of trust. To react to someone as a moral agent is to come to trust them to some extent. It is to think of the exercise of self-control as available to them.

When someone you trust to a certain extent causes harm despite the fact that they have a decisive reason not to do so and have no excuse, what happens to the trust you have in them? Given that the amount of information available to us about people is vast, and given that the extent to which we trust someone is inherently a vague matter, the

answer is that we could either ignore this information about them altogether and proceed as before, or we could add it to our stock of information about them with a view to revising the extent to which we trust them should a certain pattern develop or threshold be reached. Accordingly, my own positive suggestion is that to hold someone responsible for causing harm is to think of them as having caused harm when they had a decisive reason not to and had no excuse, and to add that thought about them to our stock of information about them with a view to revising the extent to which we trust them. So understood, to hold someone responsible for causing harm isn't necessarily to change the way that we relate to them, but it is to put ourselves into a state of readiness to change the way in which we relate to them should further relevant information about them be forthcoming (compare Scanlon 2008: Chapter Four).

Think again about the example discussed earlier in which your competitor owns a coffee shop next to yours that goes out of business because of his poor business decisions. I suggested that it is very plausible to suppose that you hold him responsible for the harm he causes to himself, but implausible to suppose either that you desire him not to have done the things he did that caused that harm, or that you were disposed to urge him to consider his own interests and make better business decisions. But note that it is isn't implausible at all to suppose that you hold him responsible because you think of him as having caused that harm to himself when he had a decisive reason not to and had no excuse, and because you add that information to your stock of information about him with a view to revising the extent to which you trust him should a pattern emerge or a threshold be reached. Imagine that he subsequently came to you with what might otherwise seem to be a promising business proposition. Depending on what happens in

the interim, you might well have misgivings precisely because you hold him responsible for the earlier failure of his coffee shop business.

I also suggested earlier that Watson is surely right that when we hold someone responsible for causing harm, we are typically disposed to engage in various forms of moral address. The account proposed of what it is to hold someone responsible explains why this is so. Imagine that we add a thought about someone causing harm like that described to our stock of information about him. If it is important that we can maintain our level of trust in him, which may well be the typical case given how much we depend on each other, then we will of course be disposed to demand that he shows consideration for those whose interests are at stake. The possibility of social life requires that, typically, we are prepared to stand up for ourselves and others against our oppressors; that we confront our oppressors with the fact of their misconduct, urging and even demanding consideration for ourselves and others. But though typical, this connection with moral address is not constitutive.

Finally, I also suggested earlier that an analysis of what it is hold someone responsible should be compatible with those who hold one another responsible having a variety of views about what the demands of morality are. Think again about Strawson who, being a retributivist, desires that those he holds responsible for causing harm suffer proportionately to the harm they cause. The account given of what it is to hold someone responsible is consistent with Strawson's holding those people responsible. Strawson surely does add a thought about the harm that those people cause without an excuse to his stock of information about them with a view to revising the extent to which he trusts

them. Watson is therefore right that the retributive sentiments can in this way be stripped away from holding responsible and the demands and reactions in which it consists.

Is the account proposed Strawsonian? The answer is yes and no. As I have already said, the account proposed of what it is to hold someone responsible is Watsonian to the extent that it takes that attitude to be a reaction to another person as a moral agent. If at the most general level this is what Strawson had in mind in suggesting that to hold someone responsible is a matter of having certain reactive attitudes towards them, then the account proposed is, to that extent, Strawsonian too. However the account proposed of what it is to hold someone responsible is not Strawsonian in another respect. It is not offered as part of an expressive theory of what it is for someone to be responsible. Someone is responsible for causing harm, I have suggested, just in case they caused that harm when they had a decisive reason not to, and they had but failed to exercise the capacity to control themselves. The account proposed of what it is to hold someone responsible thus presupposes this account of what it is for someone to be responsible. The direction of explanation is the reverse of that proposed by Strawson, and it is the reverse of that proposed by Watson too.

Bibliography

Gease, Arlyss. (2016). *A Theory of Blame and Blameworthiness*. (Princeton University PhD dissertation).

McGeer, Victoria and Philip Pettit. (2015). "The Hard Problem of Responsibility" in *Oxford Studies in Agency and Responsibility*. Volume 3. Edited by David Shoemaker. (Oxford: Oxford University Press).

Mele, Alfred R. (2012). *Backsliding: Understanding Weakness of Will*. (New York: Oxford University Press).

Rosen, Gideon. (2015). "The Alethic Conception of Moral Responsibility" in *The Nature of Moral Responsibility: New Essays*. Edited by Randolph Clarke, Michael McKenna, and Angela M. Smith. (New York: Oxford University Press).

Scanlon, Thomas M. (2008). *Moral Dimensions: Meaning, Permissibility, Blame*. (Cambridge, MA: Belknap Press).

Sher, George. (2006). *In Praise of Blame*. (New York: Oxford University Press).

Smith, Michael. (2003). "Rational Capacities" in *Weakness of Will and Varieties of Practical Irrationality*. Edited by Sarah Stroud and Christine Tappolet. (Oxford: Oxford University Press). Pp.17-38.

Strawson, Peter. (1962). "Freedom and Resentment". *Proceedings of the British Academy*. Reprinted in Watson 2003. Pp.72-93.

Watson, Gary. (1977). "Skepticism about Weakness of Will". Reprinted in Watson 2004.

_____ (1987). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme". Reprinted in Watson 2004.

_____ (2003). *Free Will, Second Edition*. (New York: Oxford University Press).

_____ (2004). *Agency and Answerability: Selected Essays*. (Oxford: Clarendon Press).