

## ACCOUNTABILITY

By Jonathan Bennett

### 1. Introduction

I shall present a problem about accountability, and its solution by Strawson's 'Freedom and Resentment'.<sup>1</sup> Some readers of this don't see it as a profound contribution to moral philosophy, and I want to help them. It may be helpful to follow up Strawson's gracefully written discussion with a more staccato presentation. My treatment will also be angled somewhat differently from his, so that its lights and shadows will fall with a certain difference, which may make it serviceable even to the converted. Also, I shall point to some disputable things in 'Freedom and Resentment', and offer repairs.

So I wrote in the first published version of this paper.<sup>2</sup> I wanted not only to be useful to others but also to elicit Strawson's certificate of approval; and that hope was realized. In his 'Reply' Strawson wrote: 'Bennett in the first eleven sections of his essay sets out and elaborates the essence of my position with such thorough and sympathetic understanding as to leave me little to say.'<sup>3</sup>

I also tried, unsuccessfully, to analyse with more precision Strawson's concept of *reactive attitude*, and to explore the extent of and reasons for the incompatibility between reactive attitudes and the objective attitude. I hoped that the display of my failures would induce Strawson to tackle the problems himself, with more success. No such luck! He wrote: 'Bennett seeks . . . to produce a tighter and more unified organisation of the phenomena . . . than I achieved in "Freedom and Resentment"',<sup>4</sup> but he did not return to the fray. On the contrary: 'It does not seem to me to matter if a strict definition [of 'reactive'] is not to be had'; and he said nothing about reasons for the reactive/objective conflict.

In the present version of the paper, I expound 'Freedom and Resentment' much as before. Since my attempts to tighten and deepen the theory failed to hook Strawson, and are not of much intrinsic interest, I now omit them. I shall, however, add an application of the doctrines of 'Freedom and Resentment' to the most basic philosophical question regarding punishment.

### 2. Accountability

We welcome some events, and regret some. Among the kinds of events that may be welcomed or regretted are human actions. When we regret an action, we may blame the agent for it, resent his doing it, hold it against him, find fault with him, speak of or to him in a manner that is censorious or vilifying or abusive, seek revenge, demand punishment. These responses are all related to *blame* - not as a faulty compass may be blamed for an accident, but in the stronger sense in which the object of blame must be believed to be personal, and the attribution of blame is a censure or reproach, which could naturally carry with it thoughts about moral unworthiness. When we

---

<sup>1</sup> P. F. Strawson, 'Freedom and Resentment', first published in 1962 and reprinted in Gary Watson, ed., *Free Will* (Oxford: Oxford University Press, 1982), pp. 59-80; second edition 2003. My page references will be to ??.

<sup>2</sup> Jonathan Bennett, 'Accountability', in Z. Van Straaten (ed.), *Philosophical Subjects: Essays Presented to P. F. Strawson* (Oxford: Clarendon Press, 1980), pp. 14-47.

<sup>3</sup> P. F. Strawson, 'Replies', *ibid.*, pp. 260-96, at p. 264.

<sup>4</sup> *Ibid.*, p. 266.

welcome an action, we may respond with praise, admiration, gratitude, thoughts of reward, or the like. These responses can be thought of as praise-related - not as one might praise a fine physique or beautiful hair, but rather as one might accompany praise with thoughts of moral worth.

I shall assume that blame-related and praise-related responses to actions constitute two classes of responses, each of which is unified enough for philosophy to be done about it. (If they do not, this essay will collapse; but so also will much of the literature.) Any unclarity about the borderlines of the two classes will be harmless, because my main points can be made in terms of responses that are well in from the boundaries.

Certain discoveries about why a regretted action was performed will lead any civilized person to regard blame-related responses as inappropriate. Suppose someone commits a murder, and it turns out that he had a brain tumour which had a crucial role in the causation of the murderous act: every victim of such a tumour would be virtually certain to commit hostile and violent acts, and this man will become a mild and reliably law-abiding citizen once his tumour is removed. In that case, it would be inappropriate to respond to the murder with reproaches etc., or to seek revenge or demand punishment. This man is not blameworthy.

Similarly, if a welcome action is explicable in a certain way, praise-related responses are inappropriate. If a benefactor was manifesting an insane compulsion to give things away, the beneficiary may welcome the gift but should not be grateful for it. This person is not praiseworthy.

It is widely believed, I think rightly, that what stops the performer of a regretted action from being blameworthy is just what stops the performer of a welcomed action from being praiseworthy. Anyway, the two sets of conditions have a large overlap, which is my topic.

By ‘accountable’ I shall mean ‘blameworthy or praiseworthy’: someone is ‘accountable’ for an action, in my usage, if a blame- or praise-related response to the action would not be inappropriate. And my concern is with a problem about the conditions for ‘accountability’ in this sense. Since appropriateness can be a matter of degree, so can accountability; but I shall mainly use all-or-nothing formulations, to keep things simple. What about actions that are not notably regrettable or welcomable? Their authors are not ‘accountable’ for them in my sense, which seems wrong. It does not matter. My concern is with cases where the question of accountability *comes up* in a natural way, namely ones that are regretted or welcomed, so that praise or blame may be appropriate.

### 3. A problem

I cannot pretend to tell, even abstractly, the whole story about the conditions for accountability; but two of them suffice for my purposes. For someone to be accountable for an action, (1) the action must relate in a certain way to his decisions, and (2) his decision-making capacities must satisfy a certain condition. More specifically: he is accountable for doing A only if (1) he would not have done A if he had decided (or chosen or willed or wanted, etc.) not to do A; and (2) he could have decided not to do A. There is no special problem about 1: if it is not satisfied, then something prevents him from not doing A, or makes him do A; and it is unproblematic to accept that the agent is then not accountable for doing A. But condition 2, and especially its use of ‘could have’, raises a problem which I shall now present as abstractly as possible.

If determinism is true, then for any event E occurring at time T there obtained at an earlier time a state of affairs that causally sufficed for E’s occurring at T and therefore causally ruled out E’s not occurring at T. Now, it is not obviously absurd to think that ‘An earlier state of affairs

obtained which causally ruled out the non-occurrence of E' entails 'E could not have not occurred'; so it is not obviously absurd to think that determinism implies that nothing which did happen could have not happened, and thus implies that there is no accountability.

There are two ways of meeting this difficulty.

#### 4. The libertarian answer

One is to suppose that determinism is false. So it probably is, in which case some events 'could have' not happened, in the sense of not being preceded by causally sufficient conditions for their occurrence. But this does not help to rescue accountability; for a chance event, whose occurrence is a matter of absolutely brute, inexplicable fact, is one for which obviously nobody is accountable. I am assuming that accountability requires intelligibility, and that something which is not caused cannot be rendered intelligible or removed from the 'brute fact' category. Causal explanation is not the only kind; but no explanation is possible for an event for which there is no causal explanation. For arguments directly in support of this, see Hobart's classic paper.<sup>5</sup> Indirect support comes from the plausibility of current philosophical theories that give the concept of cause a primary place in memory, personhood, action, and so on.

It is also instructive to look at actual attempts to base accountability on the falsity of determinism. For example, C. A. Campbell says that accountability belongs only to actions that arise partly from 'effort of will'; causes may limit what a man can do, but he is accountable only if causes leave open the question of whether the man will follow his baser desires or rather the call of 'duty', and this depends upon how much 'effort of will' he exerts towards doing what he conceives to be 'his duty'.<sup>6</sup> But Campbellian 'effort of will' cannot be what we ordinarily describe in terms of 'effort' or 'trying' or 'struggling to do one's best' or the like; for those expressions name phenomena with thick causal roots running back into the past. How hard someone tries to do his best can be affected by parental and other influences, and by his having resolved to do his best, meditated on the importance of morality, and so on. But such influences cannot bear upon Campbellian 'effort of will', which is a pure repository of uncaused determinants of action. So we don't know what Campbellian 'effort of will' is, and hence can have no reason to connect it in any way with accountability. I predict the same fate for any attempt to base accountability on uncaused inputs into action.

#### 5. The reconciling answer

The other way of trying to rescue accountability is by arguing that any sense of 'could have' which makes it true that

(1) If determinism is true, then it is never true of something that did not happen that it could have happened

is stronger than any sense of 'could have' which makes it true that

(2) An agent is accountable for an action only if he could have decided not to perform it.

I accept that there is this ambiguity, and that the non-existence of accountability therefore does not follow from determinism; and this is the currently most popular view of the matter. But it

---

<sup>5</sup> R. E. Hobart, 'Free Will as Involving Determination and Inconceivable Without It', *Mind* 43 (1934), pp. 1-27.

<sup>6</sup> C. A. Campbell, 'In Defence of Free Will', published in 1938 and reprinted in *In Defence of Free Will and Other Philosophical Essays* (London: ??, 1967), pp. 35-55; see especially pp. 42-5.

needs to be stiffened by an account of what the sense of 'could have' is that renders 2 true. The various attempts to explain this sense have in common something like this: An agent 'could have' decided differently if there was no impediment or obstacle to his doing so; the idea being that just as there are outer obstacles to executing one's will, there are also inner obstacles or impediments to the exercise of one's will. (It would be easy to restate this whole discussion in terms of 'compulsion' or 'coercion' of the will.) Now, of course, the problem has been relocated, and we must explain what an 'obstacle' is. There is no special problem about (outer) obstacles to the executing of one's will: they are just states of affairs which bring it about that one does not refrain from performing an A even if one chooses, decides, wants, etc. not to perform an A; and it is not puzzling that accountability should require the absence of such obstacles, for the blame- and praise-related responses are essentially directed to what agents willingly do.

But there is a double problem about obstacles to the exercise of the will. The item in question does not stand in opposition to the agent's will - so why is it an 'obstacle'? The action in question is performed willingly - so why is the person not accountable for it?

The former question might be answered by adducing facts about usage, exhibiting semantic principles according to which a brain tumour may count as an 'obstacle' while certain other causally sufficient conditions do not. But that will leave untouched the harder, deeper half of the problem: if tumours count as 'obstacles' or 'impediments' whereas some brain-structures do not - or if brainwashing counts as 'coercive' whereas normal education does not - then why does accountability depend upon facts about impediments or obstacles or coercion?

## 6. The Schlickian rationale

Until 'Freedom and Resentment' appeared, the only answer to this question that the literature contained was the one offered by theories of accountability like Moritz Schlick's.<sup>7</sup> These focus on the notion of moral pressure, considered as a means for changing the likelihood that the person concerned (or others who know of the 'pressure') will act similarly on later comparable occasions. Moral pressures extend from faint expressions of (dis)approval through to dire punishments and munificent rewards - a mixed bag, but all capable of generating threats or inducements, i.e. of equipping someone with a thought of the form 'If I perform an A, the upshot is likely to be U', which may affect his decision whether to perform an A. (For brevity, I shall concentrate on deterrence, ignoring encouragement; and on deterrence of the protagonist, ignoring onlookers.) Now, if someone performs an A and would not have been deterred by threats, then on future closely similar occasions he will again be undeterred by them; and so it is not useful to apply them to him by morally pressuring him in respect of the A which he has performed. That, according to Schlickian theories, explains the extent of the concept of accountability: the man with the brain tumour, for example, is not accountable for what he did because it is so unlikely that the exerting of moral pressures would have deterred him from doing it. Similarly with the person who has been brainwashed. Quite generally - say the Schlickian theories - we do not hold babies, the insane, the intellectually handicapped, the tortured, accountable for bad things they do, because it is not

---

<sup>7</sup> Moritz Schlick, *The Problems of Ethics* (trans. D. Rynin, New York: ??, 1939), ch. 7. See also Hobart, op. cit., pp. 24-7; P. H. Nowell-Smith, *Ethics* (London: ??, 1954), pp. 300-6; John Hospers, 'What Means this Freedom?', in Sidney Hook (ed.), *Determinism and Freedom in the Age of Modern Science* (New York: ??, 1958), pp. 113-30, at pp. 115-19.

useful to apply moral pressures to someone who is too young, too ill, too stupid, too hard-pressed, to be affected by them.

Accountability is strongly correlated with susceptibility to moral pressures, and so Schlickian theories draw the line - or locate the continuum - in about the right place. And they do not employ an unexplained 'could have', but only the relatively plain 'would have' which occurs in the form 'If he had thought . . . , he would have decided . . .'. Furthermore, they describe accountability in a way that offers to explain *why* the concept's limits lie where they do. Without that explanatory component, Schlickian theories would not be seriously interesting.

## 7. Why Schlickian theories are unacceptable

With it, however, they are in trouble: the Schlickian description of what accountability is - or of what the concept is for - is obviously incomplete and strikes most people as positively wrong. The latter will say that although a distinction based on the utility of a certain sort of therapy or behaviour-control might coincide with accountability/non-accountability, it cannot give the latter's essence, and that the Schlickian rationale for the line misrepresents the real nature of our praise- and blame-related responses. When we express indignation for someone's cruelty, or admiration for his unselfishness, we usually are not engaged in any sort of therapy. Blame-related responses all involve something like hostility towards the subject; whereas a moral-pressure therapist, though he may have to feign ill-feeling for therapeutic purposes, can in fact be in a perfectly sunlit frame of mind. And - to move briefly to the 'welcome' side of the fence - one may apply moral pressures to encourage a welcomed kind of behaviour while remaining in an ice-cold frame of mind, with no feelings of gratitude, admiration or the like.

Schlickians defend their omission, arguing that we ought to jettison blame-related responses and handle ill-doers purely with a view to producing the best possible outcome. But what about the praise-related responses? Schlickians never say that we should give up admiration and gratitude and settle for 'therapies' aimed at encouraging recurrences of the welcomed kind of behaviour; but shouldn't they say just that? If blame-related responses are condemned just because we cannot explain the extent of accountability except by relating it to the relevance of a certain kind of 'therapy', then there is a strictly analogous case against the praise-related responses; but obviously we ought not to give up admiration and gratitude. So something has gone wrong.

We need to make room for at least part of what Schlickian theories omit, doing so in a manner that is not embarrassed by renewed difficulties over explaining why the line falls where it does. This double need is, in my view, satisfied by 'Freedom and Resentment'.

## 8. Reactive feelings

According to Strawson, all that is omitted by Schlickian theories is the element of what he calls *reactive attitudes*, from which I shall at first lift out the component notion of a reactive *feeling*. Reactive feelings are ones that are prominent in blame, reproach, vilification, resentment, admiration, gratitude, praise and so on. (If I could define 'reactive' I would do so, rather than resorting to examples.) Clearly, Schlickian theories offer us a way of handling accountability, or some notion coextensive with it, in a manner that does not demand reactive feelings. It is a manner that does demand the *objective attitude* towards the person concerned. The phrase 'the objective attitude' is Strawson's, and the core of its meaning seems to be this: To adopt the

objective attitude towards something is to inquire into how it is structured and/or how it functions.

Many people find that feelings such as those of resentment and gratitude, indignation and admiration, do not easily occupy the mind along with a thoroughgoing concern to study the subject's behaviour patterns. That is why, as Strawson points out, one can dispel a hostile reactive feeling by cultivating objectivity of attitude towards the offender, e.g. dispelling indignation by viewing him as 'a case'.

In so far as reactive feelings won't mix with through objectivity, to that extent we must choose: we cannot always proceed as Schlick would have us do while also throwing in reactivity for good measure. Now, really, all that Schlickian theory advocates is that we let our response to each welcomed or regretted action be guided by a concern for achieving the best over-all outcome. So if reactive feelings are to have a place in our lives, we cannot always ask ourselves 'What response to that action will be for the best in the long run?' Displays of indignation or of gratitude often produce good results; but such feelings cannot be motivated by the desire to produce good results, nor, it seems, are we able closely to control them by thoughts of what will bring the best results. So apparently reactive feelings can have a considerable place in our lives only at the risk of our sometimes not acting in the most fortunate manner; and that fact might be used in a Schlickian counter-attack. Strawson's defence is to maintain that it would be unfortunate if we were always guided by the thought of what would be most fortunate: the prospect of human life with continual Schlickian preoccupations and no reactive feelings, he says, is barely conceivable and wholly repellent. And so the practical question 'Should we try to rid ourselves of reactive feelings?' is given a suitably practical answer.

### **9. The extent of accountability: Strawson's rationale**

When should reactive feelings occur? Well, to start with, Strawson marks off the area in which they would be flatly inappropriate, as fear is in the absence of danger. We could call this the area of non-accountability. Strawson gives a non-Schlickian rationale for that line's falling where it does, i.e. for the *extent* of the concept of accountability (??pp. 7-9). It has two parts, corresponding to the two main things Strawson says about the role of reactive feelings in our lives. (1) They get their value from their role in normal, adult, interpersonal relations; and so it is inappropriate to have such feelings towards someone whose youth, mental ill-health, etc. incapacitates him - whether temporarily or permanently - for such relations. (2) They are essentially expressions of one's caring about the attitudes of other people; and so they can be inappropriate because 'he didn't realize . . .', 'they couldn't help . . .', 'she didn't mean . . .' etc., where behaviour does not really manifest attitudes which it superficially seems to manifest. Under both 1 and 2, incidentally, there is plenty of room for accountability to be a matter of degree.

Without denying that this is a therapeutically useful place to draw the line, therefore, Strawson can still maintain that utility is not the whole story since he has also a non-Schlickian explanation for the line's falling where it does.

### **10. The topology of blame**

Within the area where reactive feelings are never flatly inappropriate, there may be sub-areas where they would predictably be so harmful that they should be systematically excluded from them: Strawson gives the example of the feelings of a psycho-analyst towards a patient. (Another

plausible candidate is the applying of the penal code - a topic I shall discuss in my final sections.) However, to try always to keep reactive feelings within the bounds of prudence - avoiding every counter-productive fit of pique or surge of love - would involve keeping them continuously under objective-teleological control; and that seems to be impossible. If our lives are to have a measure of warmth and engagement and spontaneity, we must pay the price of sometimes not acting in the most prudent or fortunate way. This throws a new light on the conspicuous unruliness of our emotional lives. It is not merely true, but inevitable and acceptable, that the detailed facts about when a given person has feelings of indignation, admiration, resentment, gratitude etc. partly reflect individual temperament, personal style, the mood of the moment, perhaps physiological accident.

A picture might help. There is a large 'accountability' circle within which reactive feelings are confined: it roughly coincides with the circle within which moral-pressure therapies have some chance of success. Inside that, there are smaller circles marking areas from which it is prudent to exclude reactive feelings because they are so counter-productive there. Their fully permissible range, then, consists of the area that lies inside the large circle and outside the smaller ones. They are free to roam through that area without further confinement: one cannot mark off further sub-areas within which reactive feelings are mandatory, or establish any rules of the form: if . . . , then it is wrong not to be indignant (grateful, resentful, etc.). In speaking of the ability to dispel a reactive feeling on a given occasion by cultivating thorough objectivity towards a person in question, Strawson says that we 'sometimes' have this option (?? p. 9); but I think he would and should allow that that option is always theoretically open, i.e. that it is never just wrong - though it is sometimes psychologically impossible - to dispel one's reactive feelings by retreating into objectivity. We can regard someone as 'a case' without believing or pretending that he is mentally ill etc.; for it is just a matter of thoroughly viewing him, in a spirit of inquiry, as a natural object, and this can never be 'wrong in the nature of the case'.

This has an important upshot, which for brevity's sake I shall state only in terms of blame. Consider the proposition that someone 'is blameworthy'. Strawson has a sense for this if it means that it would *not be wrong to* blame the person, but not if it means that it would *be wrong not to* blame him. There is a way of thinking about accountability according to which a person's being 'to blame' implies that blame ought not to withheld from him (though he may be spared its consequences because of forgiveness); but Strawson's account has nothing like this - no imperatives demanding indignation or any other reactive feeling, but only imperatives forbidding them in certain areas, and permissions to have them in the remaining areas.

This is one mark of the non-propositional nature of blaming, praising etc. in Strawson's account: feelings are made central, and are not tied systematically to any propositions about their objects. My feeling of indignation at what you have done is not a perception of your objective blameworthiness, nor is it demanded of me by such a perception. It expresses my emotional make-up, rather than reflecting my ability to recognize a blame-meriting person when I see one. The gap left by the Schlickian account is not to be filled by facts about desert or about the meriting of blame, facts that are acknowledged by the adoption of reactive attitudes; rather, in Strawson's words, 'it is just these attitudes themselves which fill the gap' (p. ?? 23).

Strawson says that his theory provides a basis for an understanding 'of what we mean, i.e. of *all* we mean, when, speaking the language of morals, we speak of desert, responsibility, guilt, condemnation, and justice' (p. ??23). I believe that his theory is more revisionary - or, rather, excisionary - than this implies, because I think that many people have a notion of accountability

which incorporates the belief that desert or blameworthiness or accountability is strictly a matter of objective fact. Perhaps Strawson means to claim only to have provided for every *coherent* element in ‘what we mean’, so that what is offered is not a fully conservative theory but rather a maximal salvage. That claim would be correct, I think, but I cannot prove it is. I don’t anticipate anyone’s denying that Strawson’s account is all right as far as it goes, but some may maintain that it is not the whole story. We can evaluate that claim when they tell us what the rest of the story is supposed to be.

### 11. An impasse explained

This work of Strawson’s yields benefits that he does not explicitly point out. One is a satisfying way of settling (at last!) the old issue about determinism as a threat to accountability. Many careful and intelligent people are influenced by lines of thought in which a person is presented as a natural object whose structure and behaviour ultimately results from nothing but the behaviour of parts of the universe other than himself; and in which his behaviour is presented as wholly predictable. Such lines of thought lead many people to say that the person is not really accountable for what he does - that his behaviour results from his structure and his environment, both of which are ultimately hands that were dealt to him by God or nature, so that neither they nor anything resulting purely from them should be blamed upon the person himself. I call this position ‘Spinozism’.

It is often said that Spinozism is a pure product of conceptual muddle: someone who thinks that determinism rules out accountability must be failing to grasp which sufficient conditions count as ‘compelling’, which relations to the universe count as ‘victimhood’, which sort of predictability defeats accountability, and so on. For example: ‘We have here . . . a persistent, an age-long deadlock due solely to the indisposition of the human mind to look closely into the meaning of its terms.’<sup>8</sup>

This cannot be the whole story; for many people, without being in the least muddled, hold that if a person is as God or nature made him, and if how he is determines what he does, then it is ‘in some ultimate sense hideously unfair’ that he should be blamed for bad things that he does. That phrase comes from Bernard Williams.<sup>9</sup> In the course of sketching Kant’s theory of freedom, he offers something that could explain the power of determinism to create doubts about accountability. It is that one element in ‘moral ideas influenced by Christianity’ is the thought ‘that moral worth must be separated from any natural advantage whatsoever’, a thought that led Kant ‘to the conclusion that the source of moral thought and action must be located outside the empirically conditioned self’.

One version of this frame of mind depends on the belief in a God who is ultimately responsible for every fact about the natural realm, and is also the arbiter and punisher of wrongdoing. This implies that there is something repellent about the idea of someone’s being blameworthy for an action that is an inevitable consequence of earlier states of the universe. The God of Christianity, it seems, cannot justly blame us for anything unless he has given us some kind of agency that takes our actions right out of his field of operations. The causal order he has

---

<sup>8</sup> Hobart, op. cit., p. 107 [??of reprint in his collection?].

<sup>9</sup> Bernard Williams, ‘Morality and the Emotions’, reprinted in his *Problems of the Self* (Cambridge: Cambridge University Press, 1973), pp. 207-29, at p. 228.

imposed on the universe must be incomplete, and we must be able to determine some of what happens in the gaps.

That, however, does not meet the need. When *indeterminism* is taken seriously, it seems equally at odds with accountability. When we view a human action as *not* deterministically caused, so that the totality of its causal antecedents did not settle whether the person would act thus rather than so, his acting thus strikes us as random, a matter of luck, and our sense of him as possibly to blame for his behavior is again weakened. Some will say that this is because of another conceptual muddle: we have thought in terms of mere indeterminism instead of in terms of *agent causation*. How the person acted was not fully determined by antecedent states of the universe because it was partly determined by *him*. In the absence of an account of agent causation - I mean one that is coherent, detailed, and deep - this sounds like whistling in the dark. I do not think that many people would be attracted to it if they did not see it as their best chance of rescuing accountability. I shall offer something better.

First, I should mention the not unpopular view that both sides of the impasse are right: someone's being accountable for an action is incompatible *both* with its having been deterministically caused by antecedent states of the universe *and also* with its not having been so determined; from which it follows that our concept of accountability is inconsistent, making demands that the world could not possibly meet. This might be correct; I do not hold on principle that everything must be all right with our conceptual scheme. But it would be unphilosophical to leave it at that. If we have a logically unsatisfiable concept of accountability, why do we have it? It must be because we are pulled two ways; and we should ask what does the pulling. Strawson's work could enable us to answer that question; it could let us strengthen and complete the inconsistent-concept diagnosis of the impasse, by explaining what led us into that conceptual mishap. I shall present it, however, as doing something different, namely explaining the impasse without supposing any inconsistent concept to be involved. It matters little which of these we adopt.

The Strawsonian explanation for the impasse goes as follows. When we contemplate someone's action as the upshot of deterministic causes, we adopt the objective attitude towards him; our frame of mind encourages questions like 'What do we have here? How did this come about?' which naturally goes with the question 'How can we lessen (or increase) the chance that this will happen again?' That objectivity of attitude *dispels* reactive feelings, and their disappearance presents itself to us as the judgment that the person is not morally accountable.

When instead we contemplate the action as not arising inevitably from antecedent events, we again adopt an objective attitude towards him; we are again in the 'What do we have here?' frame of mind; and so again we are pushed out of reactive attitudes towards the person in respect of this action, and we think that this has involved our giving up the judgment that he is morally accountable.

What seemed to be this:

The proposition *that P* conflicts with the attribution of moral accountability, and so does the proposition *that not-P*,

from which we might infer that the concept of accountability cannot be satisfied, is really this:

By actively raising the question '*P or not-P?*' - i.e. by thinking objectively about the action - we get into a frame of mind in which we cannot have reactive feelings; and their absence makes us reluctant to describe or treat the person as morally accountable.

Rather than moral accountability's being *logically* inconsistent with each *answer* to the question, reactive feelings are *psychologically* immiscible with the frame of mind in which the question is *asked*. The answer does not matter: the objectivity of attitude that frames the question does the real work. Dostoyevsky described it memorably:

But what can I do if I don't even feel resentment? . . . My anger, in consequence of the damned laws of consciousness, is subject to chemical decomposition. As you look, its object vanishes into thin air, its reasons evaporate, the offender is nowhere to be found, the affront ceases to be an offence and becomes destiny, something like toothache, for which nobody is to blame.<sup>10</sup>

The affront ceases to be an offence, not because of what you find when you look but just because *you look*.

In what follows, I shall use the phrase 'naturalistic thoughts about x' to mean 'intense thoughts about the causes, whether deterministic or not, of x's behaviour'.

## 12. Dispelling and disqualifying

The foregoing explanation requires (1) that when you are drawn by naturalistic thoughts about someone's actions towards the conclusion that he is not to blame for them, you are losing your feelings of indignation, etc.; and (2) that in such cases your feelings are being dispelled without being disqualified or shown to be inappropriate. Of these 1 seems clearly to be true: it fits what happens when people are swayed by Hospers's eloquent and persuasive Spinozist attack on accountability;<sup>11</sup> and I cannot imagine anyone thinking hard about the causation of behaviour while continuing to boil with rage against the malefactor. As for 2: well, it would be absurd to accept that the feelings are being disqualified, in the absence of any account of how or why.

Admittedly, I cannot explain how an intellectual operation can dispel a feeling without disqualifying it; but that gap in my position still leaves me with a reason for saying that naturalistic thoughts do not disqualify indignation, etc. If they did, the conclusion must be that if men are fit subjects of such thoughts, then we oughtn't to have reactive feelings towards them. But they *are* fit subjects for such thoughts - *everything* is - yet if we try to imagine our lives without reactive feelings we find ourselves (here I follow Strawson) confronted by a bleak desolation. We cannot be obliged to give up something whose loss would gravely worsen the human condition, and so reactive feelings cannot be made impermissible by any facts, e.g. the fact that men are natural objects about which naturalistic thoughts are possible.

That argument presupposes that the question 'Ought we to give up reactive feelings?' is a *practical* one. Anyone who construes it as such will agree with Strawson that it is to be answered 'in the light of an assessment of the gains and losses to human life' (?? p. 13); but is that the right way to construe it? It has usually been assumed that the decision about whether to permit ourselves indignation etc. must depend strictly upon whether our fellow humans are objectively (un)meritorious in some way which calls down blame or praise upon them. Even on the Strawsonian position which I am adopting, a theoretical question is involved: reactive feelings would be inappropriate if men couldn't enter into relations of love, hate, etc. But given that

---

<sup>10</sup> Fyodor Dostoyevsky, *Notes from Underground* (trans. J. Coulson, Harmondsworth: Penguin Books, 1972), p. 27 (ch. 1, section 5).

<sup>11</sup> Hospers, *op. cit.*, pp. 119-27.

conduct lies within the large circle, the remaining question is a practical one which does not strictly depend upon the establishing of any further kind of fact about the present.

Strawson does not prove this; nor can I. But the literature contains no coherent account of any relevant 'further kind of fact' and Strawson offers a liberating hypothesis which enables us to dispense with this elusive theoretical item. It is reasonable to adopt his hypothesis if it stands up while every rival falls flat.

The greatest single achievement of 'Freedom and Resentment', in my view, is its showing how the question 'Ought we to retain praise, blame, etc.?' could be a fundamentally practical one rather than having a strict dependence upon a perpetually troublesome theoretical question. Construed as practical, the question is easy to answer.

Strawson emphasizes - more than I would want to - that we could not possibly relinquish all reactive feelings. Still, ought we to try? Ought we to strive in that direction? We do have a live question about what course we should steer, and I have been expounding Strawson's answer to it.

### 13. Harmful kinds of reactive feelings

Even if we should not set ourselves against all kinds of reactive feelings, perhaps some should not be retained. Some people think it would be better if we lacked resentment and anger, etc., while retaining gratitude and every sort of love, etc.

I have heard it argued that this semi-Spinozist ideal is self-defeating, because the non-adverse reactive feelings require the adverse ones. For instance: 'You can't really love someone with whom you never get angry.' Clearly, some people hold that view of love, but the mere existence of such certainties does not count for much: men have thought that you can't really love a woman whom you never beat. I don't advocate the semi-Spinozist ideal, but I offer it as needing discussion even if one follows Strawson in rejecting the complete Spinozist ideal of relinquishing all reactive feelings.

Strawson would reject the semi-Spinozist ideal, I believe. He characterizes reactive attitudes as essentially 'participant', and associates 'sustained objectivity' with 'isolation'; and this suggests that the semi-Spinozist ideal would involve our participating in personal relations only while they please us, and withdrawing into 'isolation' whenever others behave in ways we regret. Put like that, it sounds unattractive; but I reject this formulation of the issue because reactive attitudes should not be allowed to claim the whole territory of 'participant' relationships. A therapist and her client can be closely *involved* with one another, in a therapeutic programme in which they both *participate*; but that involvement might be untouched - at least on the therapist's side - by anything Strawson would call 'reactive'.

Some people may think that we should at least try to relinquish adverse reactive feelings about ourselves, especially guilt and remorse. This could be called the Yeatsian ideal:

I am content to follow to its source  
 Every event in action or in thought;  
 Measure the lot; forgive myself the lot!  
 When such as I cast out remorse  
 So great a sweetness flows into the breast  
 We must laugh and we must sing,  
 We are blest by everything,  
 And everything we look upon is blest.<sup>12</sup>

What is in question in the third line is not the ‘reactive’ kind of forgiveness that Strawson talks about (?? p. 6), but rather the kind that consists in opting out of blame and into objectivity of attitude, into ‘measuring the lot’ - the sort that supplies whatever truth there is in *Tout comprendre, c’est tout pardonner*. Yeats’s rejection of remorse was part of his fight against ‘emotions . . . in which there is not an athletic joy’; this was not a rejection of all adverse reactive feelings, for ‘indignation is a kind of joy’. Remorse, he rightly thought, isn’t.

This Yeatsian (or demi-semi-Spinozist) ideal need not make us complacent about our past wrongdoings and failures: complacency can be warded off by self-criticism, which is consistent with perfect objectivity of attitude. To be self-critical and self-corrective, we need standards by which to judge our behaviour; but neither the Yeatsian ideal nor the all-in Spinozist one offers the slightest impediment to our judging some actions to be good or right or successful and others to be bad or wrong or failures. Without having any tendency to remorse or guilt, I may resolve not to harm other people, and when I do harm someone I may regret this very much, and be concerned to find out what went wrong - ‘measure the lot’ - and correct it.

I don’t endorse the Yeatsian ideal either; but like the semi-Spinozist one it is worth thinking about.

#### 14. The other two categories

Strawson’s account starts with ‘personal’ reactive attitudes and then adds to them self-reactive attitudes and impersonal or moral ones. Let us examine these two additions.

First, I should say more about how personal reactive attitudes are introduced. Having instanced some of ‘the many different kinds of relationship which we can have with other people’, Strawson remarks that ‘in general, we demand some degree of goodwill or regard on the part of those who stand in these relationships to us’ (?? p. 6); and he represents personal reactive attitudes as essentially a person’s response to the goodwill, indifference, etc. of those with whom he is suitably interrelated. These two elements - kinds of relationship, and demands for goodwill within them - generate the two parts of Strawson’s line around accountability: the agent is not to blame because he is incapable of entering into relationships of the relevant kind, or because his action did not really manifest a lack of goodwill. That collaboration between the two elements is a smoothly efficient affair, increasing one’s confidence that Strawson has them right.

Confidence wanes, however, when one looks at the extension of the account from personal reactive attitudes to impersonal and self-directed ones (pp. ?? 14-16). Here, Strawson makes one of the two elements do all the work. He connects personal reactive attitudes with the demand that others show goodwill towards oneself, impersonal ones with the demand that others show

---

<sup>12</sup>. W. B. Yeats, ‘A Dialogue of Self and Soul’, in his *Collected Poems* (London: ??, 1952) at p. 267.

goodwill towards men in general, and self-directed ones with the demand upon oneself that one show goodwill towards others. Nothing is said about any interpersonal relations within which such demands arise; and indeed Strawson says explicitly that the moral reactive attitudes ‘permit . . . a certain detachment’ (p. 4) and remarks on their not needing to include ‘antecedent personal involvement’ (p. 17). The impression is conveyed that to have impersonal or self-directed reactive attitudes is just to ‘acknowledge the claims’ of men upon men or of others upon oneself. Strawson does not quite say this, but it is suggested by his silence regarding what else is involved in these two kinds of reactive attitude.

But there must be more to them than that. I might ‘acknowledge’ your ‘claim’ to my goodwill, and thus regret my failures to give it to you, yet handle these lapses through self-criticism and self-amendment with no tincture of guilt or remorse; in which case I acknowledge the claims but do not have the corresponding kind of reactive attitude (unless Strawson counts self-criticism as ‘reactive’, in which case I am lost). Analogously, I might hold strongly that people should show goodwill towards one another, yet not be indignant when they fail to do so; for I may adopt a non-reactive, clinical, corrective, objective attitude to every instance I encounter of man’s inhumanity to man. I submit that the answers to the questions ‘Why does he regret [welcome] that action?’ and ‘Is his response reactive?’ are logically independent of one another. If they are, then impersonal and self-directed reactive attitudes cannot be fully explained in terms of the acknowledging of claims: those ‘acknowledgements’ explain welcomes and regrets, but cannot explain the reactivity.

In his Reply Strawson wrote: ‘I freely admit that ‘acknowledgment of claims’ is too weak a phrase.’<sup>13</sup>

## 15. Relocating one element

Personal reactive attitudes are introduced, as I have noted, through the notions of claim-to-goodwill and interpersonal relation. Now, Strawson speaks of the goodwill that is demanded in certain relations; but isn’t it also demanded outside of them? And cannot the latter demands also generate reactive attitudes? And, to take in some of the territory not covered by ‘claim’, cannot gratitude, for instance, occur without any antecedent personal involvement? ‘But in all these cases’ - you might say - ‘there is the “involvement” created by the very behaviour to which the reactive attitude is a response.’ That is true, but Strawson is not thinking of kinds of involvement or ‘interpersonal relation’ that could be created just by kicking somebody or throwing him a coin. A large theme in ‘Freedom and Resentment’ is the contrast between the involvements that go with reactive attitudes and the ‘isolation’ that would be entailed by their absence, as well as the ‘relief from the strains of involvement’ (p. 12) that comes from replacing reactive attitudes by the objective one; and all of that is reduced to nonsense if one construes ‘involvement’ etc. so as to include mere helpings and harmings.

As for the converse: Strawson clearly implies that the fact that ill-will occurs within a relationship of the emphasized sort does not guarantee that the response to it will be reactive.

One might conclude that the notion of interpersonal relation is not supposed to help explain what a personal reactive attitude is, and is offered only as part of the natural history of reactive attitudes - a mere description of their place in our lives. But that is hard to reconcile with the

---

<sup>13</sup> Strawson, ‘Replies’, op. cit., p. 266.

amount of weight Strawson seems to lay upon such expressions as ‘participant’ and ‘non-detached’. Fortunately, there is another way out.

It is to give the notion of an interpersonal relation (of the relevant kind) a role in the analytic or explanatory part of the account, but not quite the role initially allotted to it by Strawson. What should be emphasized, I suggest, is not the relations *within* which reactive attitudes arise, but rather the relations *towards* which they point. If I resent someone’s treatment of me, there may have been antecedently no special kind of relation between us; but my very resentment creates one, or sets the stage for one. I cannot say precisely what the ‘special kind’ is: that belongs to the problem of *defining* ‘reactive’, which I kicked around in the original version of this paper and now set aside. But any attempt to solve that problem, provide that definition, should be helped by the point I am now making: the participations and involvements that Strawson emphasizes should be seen not primarily as the ground in which reactive attitudes grow but rather as embodied in or consequential upon them; not as required in the past or present, but as implied or suggested or invited for the future.

This idea, though implicit throughout most of ‘Freedom and Resentment’, needs more explicit emphasis than Strawson gives it. It could lead to a tightening of the curiously loose and structureless paragraph in which ‘reactive attitudes’ are first introduced (p. 6). It could also let us strengthen a soft spot in Strawson’s rationale for the line around the concept of accountability. The reason why ‘seeing someone [as] deranged or compulsive’ tends to ‘set him apart from normal participant reactive attitudes’ (p. 9) is that those attitudes connect with normal interpersonal relations. Connect how? If reactive attitudes essentially embody or point towards or prepare for interpersonal relations, then it is clear how someone’s incapacity for the latter makes it inappropriate to have reactive attitudes towards him. But if the connection is just that reactive attitudes (should?) arise out of events between people who are interrelated, it is not clear how the argument runs. It would apparently have to put ‘He is deranged’ on a par with ‘He is a stranger to me’; in each case there is no significant relationship between us, and so (for some still unclear reason) it would be inappropriate for me to have a reactive attitude towards him. Things go better if when reactive attitudes are seen as pointing towards possible or imagined future interpersonal relations rather than as growing out of past ones. (That view of them, incidentally, agrees with the etymological roots of ‘attitude’, which comes from the Latin *aptus* - apt or fit for a given kind of action.)

By relocating the notion of interpersonal relation in this way, we get good help with Strawson’s extension of his account to cover the other two categories of reactive attitudes. A self-reactive attitude does involve an important ‘interpersonal’ relation: remorse, for instance, can be represented as a confrontation - with an accusing glare on one side and downcast eyes on the other - between one’s present self and some past self. I offer this as a realistic view of what self-reactive attitudes are like, though admittedly a still incomplete one; and as better than an account that focuses on the acknowledging of claims.

Similarly with impersonal reactive attitudes: moral indignation - we can now say - involves actually or imaginatively putting oneself into, or readying oneself for, a special kind of relation, with the person towards whom the indignation is directed.

This in turn throws light on Strawson’s view that ‘moral’ reactive attitudes are significantly more ‘detached’ than personal ones are. In my revised version, the important kind of ‘interpersonal relation’ is equally present in both categories. Often in moral cases the attitude is only an entertaining of an imagined relation, but the same is true in many personal cases, e.g.

gratitude to a dead benefactor, anger at an unidentified thief, resentment towards an oppressor whom one hopes never to see again. There is this much in Strawson's thesis: a reactive attitude of the kind he calls 'personal' is a response to someone's attitude towards oneself, and so personal reactive attitudes must be in that sense self-involving. But I see no reason to think that they must pertain to the important kinds of interpersonal relation to a greater degree than the 'moral' ones do.

This presumably connects with Strawson's suggestion that 'the tension between objectivity of view and the moral reactive attitudes is perhaps less than the tension between objectivity of view and the personal reactive attitudes' (p. 17). I can find no reason to agree with this. (Indeed, severe objectivity seems more apt to banish a blaming attitude than a resentful one; but the evidence for this might be reinterpreted as showing that objectivity is harder to achieve where one's own interests are concerned. So that is a stand-off.) I can only conjecture that Strawson was guided here by the idea that moral reactive attitudes are significantly more 'detached' than personal ones are; so this is another burden which is lifted from our shoulders if the notion of interpersonal relations is relocated in the manner I have advocated.

The relocation may also contribute a little towards explaining why reactive attitudes will not mix with the objective attitude. (Some people claim that the two can cohabit in *their* minds, where reactivity persists even when they 'look' with Dostoyevsky or 'measure the lot' with Yeats. This conflict of testimony could reflect our unclarity as to what the issue is; but the phenomenon may be subject to real interpersonal variation. Still, something needs explaining here.) Strawson does not discuss why there should be any conflict or tension, merely emphasizing how *different* the objective attitude is from reactive ones; but why does that set them against one another? Is it just a matter of the limits on how much mental variety one can manage at a single time - limits which a virtuoso of the inner life might transcend through practice? I tentatively suggest a different account. Reactive attitudes essentially prepare for personal interaction of a certain kind, while the objective attitude prepares for inquiry, and these two sorts of activity are somehow incompatible. If that is right, the two sorts of attitude are derivatively in conflict, like simultaneously readying oneself for a sexual encounter and for giving an after-dinner speech. Even if that is right, however, more work has to be done to make this matter clear.

## 16. Generalizing the other element

The relocation of the notion of interpersonal relation frees us to reconsider the demand-for-goodwill element in the account. It is salutary to be reminded of how much we care about the attitudes of others towards ourselves and towards one another; but I contend that this 'caring' belongs to the natural history of reactive attitudes rather than to the elucidation of 'reactive attitude'. We can understand the idea of someone's being genuinely morally indignant over someone else's attitude to natural beauty, for instance; and so moral indignation does not have to be a response to someone's attitude towards people. If nothing else really merits moral indignation, that is a substantive moral truth rather than a fact about the concept of moral indignation.

Essentially the same point holds for non-moral reactive attitudes, but here there is a terminological snag. Strawson assumes - rightly, in my view - that an attitude counts as 'moral' only if it rests on a general principle, or anyway on something that does not essentially refer to any particular item. So an attitude of mine is not moral if its basis essentially involves myself: I am morally indignant at your contemptuous attitude towards a benefactor, but I resent your

contemptuous attitude towards me. But the basis for an attitude might lack generality - thus depriving the attitude of the status of 'moral' - in some quite different way. For a bit of behaviour might enrage me on a particular occasion, although it neither infringes any general principle that I hold nor essentially involves myself. For instance, I take no general stand on attitudes to natural beauty, but on this one occasion it just makes me angry to see a man walk unheedingly past the masses of Alpine Lilies and Indian Paintbrush. Or I have an unreasoned 'thing' about Bruckner, which leads me to feel something like gratitude towards anyone who loves his music. That anger and the 'gratitude' are both reactive, I suppose; but they are not 'moral', since one concerns a particular occasion and the other a particular person, and neither rests on general principles. But neither of them fits comfortably under Strawson's label 'personal reactive attitude', since that label so naturally suggests an attitude which responds to someone's attitude to oneself.

I suggest, therefore, that the two basic categories of reactive attitudes are 'non-principled' and 'principled' (or 'moral'), with 'personal' as an important species within the former.

As for self-reactive attitudes: some of them are principled and some are not; for an attitude of self-censure or self-congratulation may, but need not, rest upon some principle that one holds. Strawson focuses primarily on the principled ones - which rest on one's acknowledgement of others' claims on one's goodwill - and perhaps they matter most. But there are others, such as self-reproach for having made a fool of oneself in public. Incidentally, Strawson's use of the word 'moral' is unsatisfactory on any showing, for his 'moral' category positively excludes self-reactive attitudes. Admittedly, what is 'moral' must have a general basis; but that is no obstacle to allowing that self-reactive attitudes can be moral. My remorse over my cruelty is as principled as my indignation over yours.

So much for taxonomy and terminology. Returning now to the substantive point: I suggest that although it is all right to tie reactive attitudes to responses to *somebody's attitude*, it is unduly narrowing to tie them to responses to *somebody's attitude towards somebody*. The '... towards somebody' bit looms large in the natural history - and perhaps also in the ethics - of reactive attitudes, but not in the account of what reactivity is.

When the account is thus generalized, it covers cases that are untouched by Strawson's treatment. Also, as I have shown, it forces us to make independently worthwhile revisions in the taxonomy and terminology. And, finally, it makes no difference to the relevant part of Strawson's rationale for the line around accountability. For that concerns cases where, despite appearances to the contrary, 'the agent's attitude and intentions [are] what we demand they should be' (p. 7); and this need not draw on the idea that the relevant 'attitude' must be towards one or more people.

## 17. A problem about punishment

This work of Strawson's supplies a basis for clearing up some long-standing troubles concerning the justification of punishment. I plan to explore this topic more fully elsewhere, but the pointers I give here may be of some use.

Here is a convict; how should we treat him? There are two ways of coming at an answer to this. (1) The forward-looking way says that we must consider only what treatment will maximize utility, that is, do the most over-all good. For present purposes it does not matter how that good is distributed between reforming the convict, deterring him and others, placating victims and their kin, and so on. (2) The partly backward-looking way says that we must also take into account a fact about the past, namely the fact that he did commit a crime together with the facts about how grave a crime it was.

Those who confine themselves to approach 1 - ‘utilitarians’, for short - are accused of paying no attention to guilt or innocence. When someone has been *wrongly* convicted of a crime, it may be best over-all if he is nevertheless treated in a punishing way. This offends our sense of justice, as does the related possibility that someone who is guilty of a minor crime may on utilitarian grounds be assigned a harsh punishment. (In this context, the relevance of whether there was a crime stands or falls with the relevance of how grave a crime it was. I find it helpful to think of innocence as committing a crime with gravity = 0, and to think of not being punished as receiving punishment with severity = 0.) The utilitarian can argue that it is not so easy to describe a case where utility really *would* be maximized by a deliberate injustice - penalising an innocent person, or penalising a guilty one with a harshness disproportionate to the gravity of his crime; and we can trade intuitions about cases. For most of us, though, the question of whether and how gravely the person has offended is *directly* relevant to whether and how he should be punished; which is to say that we favour approach 2.

The word *desert* and its kin naturally come in here: the facts missed by utilitarianism relate to what the person in the dock *deserves* to have done to him. This, however, does not explain much. To say that offenders deserve to be punished is to imply that their guilt is a reason - a direct, immediate reason - why it is all right to punish them. We can agree with this while still wanting to know why it is so, wanting to get this judgment from something deeper and/or more general; and the mention of ‘desert’ does not supply it.

Indeed, desert is not much of a concept: I can find almost no serious attempts to explain or analyse it, and I have never seen it being used in an argued defence for either answer to the most important question involving it:

If someone deserves to be punished in a certain way, does it follow that he *ought* to be punished in that way, even if the over-all consequences would be better if he were punished less severely or not at all?

More briefly:

Do facts about someone’s desert ever imply that he should be punished against utility? The literature contains firm affirmative answers, and equally robust negative ones. The latter come mostly from people who acknowledge that desert theorists used to answer that Yes, people ought to be punished as they deserve, even against utility, but that in these enlightened times nobody believes that any more. These pronouncements, pro and con, are never accompanied by arguments in which the concept of desert is at work - or indeed by arguments of any other kind. This is a striking manifestation of the concept’s theoretical poverty.

As a matter of morality, I take my stand with those who say not merely that we are never morally required to punish against utility but that we are morally required not to do so, but I also hold that the maximizing of utility is not the whole basis for a system of punishment. I want a *basis* for that pair of moral intuitions - something deeper and more general from which they follow. That is what can be found in Strawson’s ‘Freedom and Resentment’.

## 18. Outline of a theory of punishment

A defensible penal action does over-all good by means that bring suffering to one person. Not being outright utilitarians, we are not always willing to avail ourselves of opportunities to do this. Suppose that ceremonially hurting one randomly chosen person would be sure - through some mysterious but well-tested mechanism - to reduce the incidence of some really nasty disease. Most

of us would be reluctant to go through with the ceremony, even if we judged that that one person's suffering would be less bad than a state of affairs in which the disease is endemic.

Why? We are moved by sympathy for the one person - a fellow-feeling for someone in distress - but sympathy should have a place on the other side of the equation as well. Why does it not? Or why, at any rate, does it function differently there? Well, for most of us, harm to an identified person outweighs help - or lessened harm - to an unidentified group. Even when punishing a particular criminal will certainly bring benefits, we may be unable to point to any particular people as the beneficiaries. There may indeed *be* no people of whom it is determinately the case that *they* would benefit from this convict's being punished. This seems to affect our moral thinking. We serenely launch building projects in which probably some workers will die; compare that with how we react when a particular child is trapped in a well.

There are other possible contributors to our unwillingness to hurt one in the interests of many, though they do not do much credit to our intelligences. Anyway, we *are* in general reluctant to harm one person in the interests of many, and that reluctance seems unbudgeable. I shall take it as a given and work on from there.

Now, *punishing a criminal* should be seen, I submit, as a species of *hurting one person in the interests of many*. The moral or emotional obstacle to doing that is less in this species than throughout the rest of the genus; the fact that the one has committed a crime makes a difference. 'Because criminals deserve punishment', some will say. I agree, but I cannot make attributions of desert *explain* anything. The best answer I can find looks not to our moral principles but rather to an aspect of our nature lying deeper than our moral principles and helping to produce them and - insofar as anything can - to justify them. In blaming a convict for his behaviour, we do not assent to a proposition about him but rather adopt towards him the reactive attitude that Strawson calls 'vicarious resentment', or resentment on behalf of his victims; this essentially involves at least incipient hostility or ill-will towards its object; and this makes us less unwilling for its object - in our present case, the convict - to be hurt in the interests of the greater good. We already have indignant, offended or resentful feelings towards him, and these reduce our unwillingness that he should be made to suffer for the general good. Thus Adam Smith: 'Gratitude and resentment . . . are the sentiments which most immediately and directly prompt to reward and to punish. . . . That action must . . . surely appear to deserve punishment which everybody who hears of it is angry with, and upon that account rejoices to see punished' (1897: 286, 287).

That underlies the common idea that *it is all right to punish the guilty*. From a starting-point that differs from utilitarianism because it involves a reluctance, in many cases, to do good by inflicting suffering on one person, we introduce resentment and its vicarious analogue, blame, to lessen that reluctance and bring us closer to utilitarianism. Why is it generally wrong to punish the innocent? Because the explanation of why it is right to punish *anybody* applies only to the guilty.

This account of punishment has two desirable features. (1) It does not morally permit us - let alone require us - to punish someone more harshly than would maximize utility. The source of that feature of the account is not a mere add-on, but rather the account's driving force, namely the thesis that punishment is to be engaged in only as a way of doing good. And room is made for this by a basic fact about reactive attitudes, namely that they are sometimes permissible *and never required*. That secures that nothing in the nature of things can require us to punish someone if considerations of utility go against our doing so.

It is also a merit in this account that it provides a *direct* relevance between gravity and severity - doing this right out to the end of the scale where innocence is directly a reason for not punishing the person at all. We are initially morally reluctant to punish anyone, and for an innocent person nothing overcomes that reluctance. Similarly, nothing much overcomes it in the case of a mild offender. In this theory of punishment, innocence and degrees of severity are relevant not only because of the traces they will leave in the future but also directly.

This account puts an openly retributivist element into punishment: in it, the willingness to punish someone arises *directly* from the belief that he has committed a crime. Some defenders of retribution have taken pains to prevent it from brushing up against the idea of revenge, which they think would taint it. I have no such scruple. While I do not find it helpful to describe the crucial reactive attitude as ‘vengeful’, it certainly has the same human roots as the desire for vengeance.

Although my account is retributivist, it frames punishment within utilitarian considerations, not allowing any punishment that goes against utility. Sher writes (1987: 74): ‘To assert that we can justify punishment only by showing that it brings advantages is to beg the question against retributivism.’ Though that is plausible, I have shown it to be false. Griffin writes:

What would be entirely wrong would be to try to introduce utilitarian reasons into desert. As we have seen, it destroys a response to inject extraneous considerations into it, and utilitarian reasons are extraneous. Authenticity is not merely the best or purest form of responses such as admiration, gratitude, or appreciation; it is the only form. (Griffin 1986: 259)

That looks like a partial list of reactive attitudes, and I suppose that Griffin would include resentment in it (though he does not mention Strawson’s work). I agree that one cannot adopt a reactive attitude for a purpose: there cannot be contrived or considered or judicious resentment. But that does not place such a high barrier between punishment and utility as Griffin apparently wants to erect there. Reactive attitudes are essentially spontaneous, adopted without the guidance of a *telos*; but one can - and civilized people do - have goal-directed policies for delimiting areas of life within which they will deny themselves the luxury of reactivity.

In a more complex way, it is open for us to give play to our generalized vicarious resentment of criminal behavior by endorsing a system of punishment, expressing our resentment in our willingness to put the offender at the disposal of the legal system; while at the same time resolving that this should be enacted only in cases where it has a good enough chance of doing some good. The permissive framework is set by utility; the punishment within that framework expresses our reactive, retributive anger or resentment. Thus, a genuinely retributive element in punishment cohabits with severe utilitarian constraints.