4
P. F. Strawson's Consequentialism

Victoria McGeer

PRELIMINARIES

P. F. Strawson's 1962 address to the British Academy, "Freedom & Resentment",[1] surely constitutes one of the most remarkable and groundbreaking papers written on the topic of moral responsibility in the last half century.[2] It has certainly been one of the most influential papers: not only has it spawned an enormous secondary literature, it has provided inspiration to a wide range of philosophers who have themselves contributed in an original and substantial way to a contemporary understanding and defence of agential responsibility. Although sceptical worries about our ordinary attitudes and practices of holding responsible certainly remain

[1] All page references in this paper will be to the reprinting of "Freedom and Resentment" (hereafter, F&R) in Strawson, P. F. (1974).

[2] This paper was originally presented in September 2012 at the "Responsibility and Relationships" conference held at the College of William and Mary in celebration of the fiftieth anniversary of the publication of "Freedom and Resentment", and I thank the organizers of the conference, in particular Neal Tognazzini, for giving me the opportunity to write so directly on Strawson's views. I would also like to thank my commentator, Bennett Helm, for trying (ever so gently) to pull me back into the plainly more sensible non-consequentialist reading of Strawson's views. That he didn't succeed is no fault of his. I remain benighted and happily so, though the paper improved through my interaction with him and with other conference participants. Versions of this paper were also presented to the normative philosophy workshop at Princeton University, to the Australasian Moral Philosophy workshop in Kioloa, and to a joint session of Chapel Hill and Duke University philosophers. I am grateful for the many rich discussions I had on all these occasions; but special thanks for follow-up comments are owed to Geoff Sayre-McCord, Nate Sharadin, Walter Sinnott-Armstrong, Nic Southwood, Susan Wolf, and two anonymous referees for this volume. This paper would not have been written without the help and support of Philip Pettit, whose work on consequentialism has clearly influenced me and given me a way to articulate some long-percolating ideas about "Freedom and Resentment".

in the aftermath of Strawson's paper, there is no denying that it reset the terms of the debate, breathing new life into old ways of approaching the topic and confronting responsibility sceptics with a new range of issues to address. And yet despite its unquestionable importance in the philosophical landscape, "Freedom and Resentment" is generally acknowledged to be a puzzling work both exegetically and philosophically. Most commentators agree that, while the argument of the paper is subtle and complex, it is not entirely clear what the argument is; and beyond that, whether Strawson's view (however we understand it) is ultimately stable or satisfying. For that reason alone, it seems a good way to celebrate the fiftieth anniversary of its publication to offer an interpretation of the paper itself, focussing explicitly on what Strawson took himself to be arguing for and against. This is my aim in what follows. Still, no critical commentary would be worth its salt if it failed to make some distinctive contribution to an already substantial literature. So let me preface the interpretive work I engage in below with three further remarks on why I think this exercise is worthwhile.

First remark: As my title indicates, the position I ascribe to Strawson is avowedly consequentialist. As I read the text there are in fact *three* mutually reinforcing prongs to Strawson's defence of a robust conception of responsibility in "Freedom and Resentment": his *naturalism* (by which I mean a resistance to certain metaphysical considerations and/or debates about free will in light of our natural human attitudes and commitments); his *pragmatism* (by which I mean an emphasis on our everyday attitudes and practices of "holding responsible"); and his *consequentialism* (which I will elucidate presently). The first two prongs of Strawson's view are often recognized and discussed; the third is not. So, the interpretation I offer here is a minority view that surely deserves a hearing. At the very least, I hope it generates some lively discussion and forces a clearer articulation of why the position Strawson defends cannot be assimilated within a consequentialist framework.

Second remark: As indicated above, Strawson is commonly viewed as advocating a *non*-consequentialist approach to the problem of responsibility. This view is broadly adopted by Jonathan Bennett (1980), Gary Watson (1987), Stephen Darwall (2006), and R. Jay Wallace (1994)—to name but a very few. While such commentators are generally laudatory, emphasizing the many valuable insights to be gleaned from Strawson's approach, a common critical theme has surfaced in much of this discussion: It is that Strawson's position provides an incomplete, or inadequate, response to incompatibilists who continue to worry about the metaphysical grounding of our responsibility practices; that it lacks the resources to spell out in its own compatibilist terms a fully satisfying account of what makes someone a responsible agent; and that it needs important supplementation and/or amendment guided by extra-Strawsonian considerations—that is, considerations that take us

beyond Strawson's naturalism and/or pragmatism. I'm not surprised at this common theme. As I said above, I think there are *three* prongs to Strawson's view that are mutually reinforcing—his naturalism, his pragmatism, and his consequentialism. Together, they make up a tripod that stabilizes his position. Remove any leg of this tripod and his account of responsibility is bound to be incomplete and dissatisfying. So, on my reading of Strawson, the consequentialist prong is not an optional add-on. It lies at the very heart of his view and can't be jettisoned without leading to the kinds of problems that have been raised by a number of critics. I gesture at how Strawson's consequentialism can provide an answer to these problems in my conclusion.[3]

Third remark: An important reason for emphasizing the consequentialist aspects of Strawson's work is that the resulting position represents an independently attractive vision of what makes for moral responsibility. So my aim is not purely exegetical; it is also promotional. As I see it, a significant part of this promotional task is to defend a version of consequentialism that can take on board the critical remarks Strawson aims at the so-called "optimists" in "Freedom and Resentment". For I certainly don't deny Strawson's indictment of a certain consequentialist line of thinking about our practices of praise and blame, punishment, and reward. Yet, for all that, his project was not to toss the baby out with the bathwater, but rather—to borrow a phrase of his from elsewhere—"to strike some delicate balances" (Strawson, P. F. 1961: 8); and the balance, I will argue, is well in favour of a cleaned-up, more sophisticated version of consequentialism. That Strawson should defend such a view is no surprise to me, given its manifest advantages. But here I simply emphasize that this view is worth attending to in its own right, whether or not the exegetical claims of this paper are accepted.[4]

The argument of my paper will be in three sections. In Section 1, I outline in a very basic and familiar way the nature of the conflict, as Strawson sees it, between optimists and pessimists about what ultimately grounds and

[3] I take these problems to fall into two related categories: (i) giving an intuitively satisfying (but non-libertarian) answer to what makes someone an appropriate (or "fair") target of our reactive attitudes and practices; and (ii) finding a deeper, more adequate response to the challenge of whether we could, or should, jettison some or all of our reactive attitudes and practices.

[4] R. Jay Wallace remarks ". . . there is no fixed and stable view that might be labeled the Strawsonian account of responsibility. Strawson's original lecture contains a wealth of ideas, and many philosophers who have been influenced by the lecture have naturally chosen to develop and defend different ones among them, and to develop them in different ways" (Wallace 1994: 10). Wallace attributes this complex legacy to an original complexity in Strawson's argument, implying—perhaps—that there is no one right reading of Strawson's text. Perhaps so, but in that case I only insist that the consequentialist elements in Strawson's thought are clearly present and are well worth bringing to the fore.

justifies treating someone as a responsible agent—and why neither view is adequate. I end with a representation of Strawson's positive view that I take to be fairly standard and uncontroversial. In Section 2, I argue that this representation is right so far as it goes, but importantly incomplete. I defend a version of Strawson's positive account of justified attributions of responsibility that stresses the forward-looking *regulative* dimension of reactive attitudes. Thus, I claim, he retains an important dimension of the optimists' view, despite taking to heart the core intuitions that fuel the pessimists' flight into "panicky metaphysics". In Section 3, I turn to the question of how best to characterize Strawson's account in the larger philosophical landscape. Here I make my pitch for taking Strawson to be a kind of consequentialist, albeit of a far more sophisticated sort than are the optimists he criticizes. And, finally, in a short conclusion, I return to the question of whether any value is added by reading Strawson in this consequentialist light.

1. THE BACKGROUND DEBATE AND THE BEGINNINGS OF A RECONCILIATION

The opening pages of "Freedom and Resentment" rehearse the classic positions of a long-standing debate over the implications of the metaphysical thesis of determinism for our moral concepts and behaviour—in particular, for the key concept of moral responsibility and those of our interpersonal attitudes and practices that can be placed under the general rubric of "holding responsible". Such attitudes and practices only make sense, and can only be justified, when they are directed towards morally responsible agents. Paradigmatically, they include praise and blame, punishment and reward—though, of course, it's part of Strawson's general project to remind us of the wide array of (emotionally infused) attitudes and practices that presuppose a substantial notion of responsible agency. I'll return to Strawson's idea of "reactive" attitudes and practices in more detail in Section 2, but for now I will stick to the classic debate.

The dominant positions in this debate are represented by Strawson's so-called "optimists" and "pessimists";[5] and I here distinguish these positions in terms of three substantive issues on which they divide:

i. *Justification question*: The appropriate rationale or justification for holding people responsible—thus, paradigmatically, for the attitudes and practices involved in praise and blame.

[5] The minority voice in this debate is represented by Strawson's "genuine moral sceptics", who maintain that "the notions of moral guilt, of blame, or moral responsibility are inherently confused. . ." so far as they "lack application" no matter whether determinism

ii. *Presupposition question*: The kind of power or capacity these attitudes and practices presuppose in an agent.

iii. *Metaphysical question*: Whether this presupposition conflicts with the truth of determinism.[6]

Optimists make the following claims:

Justification question: The only acceptable rationale for engaging in the attitudes and practices of holding responsible is the *forward-looking* one of regulating behaviour in socially desirable ways. Call this the "moral pressure" view (Schlick 1939: chapter 7).[7]

Presupposition question: The moral pressure view presupposes that people will be appropriate targets of praise and blame only so far as they are rational agents, who act in accordance with their own "freely" made decisions, where these decisions are the deliberative products of rationally formed beliefs and desires. Of course, the kind of freedom involved here is best understood in a purely negative sense—freedom *from* external coercion of an unacceptable sort, and freedom *from* various sorts of internal psychological delusions or compulsions.

Metaphysical question: The truth of determinism in no way undermines this view—on the contrary, many would say it presupposes it. Agents can only be responsive to the regulative power of praise and blame, punishment and reward, so far as they form beliefs and desires that are causally sensitive to such inputs from their environment.

is true or false (F&R: 1). Such a position is interestingly opposed to the view Strawson defends in this paper—that such notions *have* application no matter whether determinism is true or false; and yet he gives these moral sceptics almost no further consideration. A resounding defence of this sort of sceptical view can be found, ironically close to home, in the work of Galen Strawson (1994).

[6] The thesis here implied is, crudely speaking, the metaphysical view that every event, including every human action, is entirely determined by the prior physical state of the universe in accordance with natural law. Strawson says, rather coyly, that he does not "know what the thesis of determinism is"; yet adds, "of course, though darkling, one has some inkling—some notion of what sort of thing is being talked about" (F&R: 1). Why this reluctance to say straightforwardly what the thesis is supposed to be? Perhaps he is driving home the lesson that, in his view, we don't need to go beyond "the facts as we know them" to resolve the debate over moral responsibility. An understanding of the thesis of determinism (which has occasioned much philosophical debate) is, in Strawson's eyes, certainly going beyond the facts as we (ordinary folk) know them. (Note: the phrase "the facts as we know them" is a significant one for Strawson and recurs throughout the text, with a first instance on p. 2.)

[7] Borrowing the term from H. L. A. Hart (1985), Wallace (1994: ch. 3) refers to this as an "economy of threats" view. T. M. Scanlon (1988: Lecture 1) uses the term "influenceability theory". Apart from Schlick, other prominent defenders of this view include J. J. C. Smart (1961), P. Nowell-Smith (1948), and (in more contemporary terms) Daniel Dennett (1984).

Pessimists shrink from this view in horror, convinced, as Strawson says, that something "vital" has been left out of the picture. This is reflected in the position they take on these three issues in turn:

Justification question: The most important difference with the optimists comes on this question. The only acceptable rationale for engaging in practices of praise and blame is backward-looking. Specifically, praise and blame should be directed towards agents only because of what they did, and only because what they did "merits" or "deserves" the response in question. Praise and blame are *earned* responses; they redound to the agents' credit or discredit—and that would be true whether or not praise or blame had any salutary effects on their future behaviour. Call this the "merit" (or "desert") view.

Presupposition question: The merit view leads to a substantial difference on this question as well. In order to truly deserve praise or blame, pessimists say, agents must be the *ultimate* source of their actions, the *ultimate* locus of control; they can't be simply a relay station for states of belief and desire over which they exercise no final say. And this implies, as Strawson says, an obscure and murky kind of "libertarian" freedom that "goes beyond the negative freedoms that the optimist concedes. It is, say, a genuinely free identification of the will with the act" (F&R: 3).

Metaphysical question: The presupposition that (truly) responsible agents exercise such libertarian freedom is incompatible with the truth of determinism.

Now the aim of "Freedom and Resentment" is to reconcile these camps—or, as Strawson says, to win "a formal withdrawal on one side in return for a substantial concession on the other" (F&R: 2). But how is this reconciliation to be achieved? As a first pass, we might reasonably characterize Strawson's view as follows, in terms of the three issues just canvassed:

First, consider the justification question: In Strawson's estimation, the substantial concession must come from the optimists, and it must be made right at the start, correcting something deeply problematic in their response to the justification question. From a phenomenological perspective, it seems indisputable that our attitudes and practices of praise and blame have precisely the backward-looking character that pessimists identify. In the normal case, when we praise and blame people, we're not engaging in a kind of behavioural therapy, thinking of how our reactions might prod them into doing the things we approve of and avoiding the things that we don't. We're reacting to what they have done—and we're reacting in a way that is permeated by the sense that they *merit* or *deserve* the good or ill reaction that we

direct at them.[8] Of course, we *can* adopt a more therapeutic or engineering approach to shaping others' behaviour—but to the extent that we do, we lose a vital feature of ordinary inter-personal relationships: the feature of demanding and showing respect for one another as morally responsible beings.

This is where the optimists go badly wrong. The "lacuna" in their position is made evident by the phenomenology of our ordinary practices of praise and blame. But the blunder they make is really a conceptual one, revealing a deep incoherence at the very heart of their story. After all, optimists agree that our attitudes of praise and blame are only properly directed towards morally responsible agents—agents that, by their own lights, have the rational capacities of autonomous beings. But so far as they recommend treating such agents as mere objects to be manipulated by the calculative application of carrots and sticks, they would have us fail to accord such agents the respect we owe them *qua* rational autonomous beings. Their recommendation thus amounts to disregarding the very status or condition that makes our praise or blame appropriate in the first place. So the optimists' position seemingly undermines itself.

On the justification question, then, it seems clear that Strawson sides with the pessimists. Efficacy in regulating behaviour cannot be the "only reason", as he says, for engaging in practices of blame and punishment. Indeed, he adds (now in the persona of the pessimist): "this is not a sufficient basis, it is not even the *right* sort of basis, for these practices as we understand them" (F&R: 4). I need hardly point out that Strawson's appreciation of this aspect of the pessimists' complaint encourages a non-consequentialist interpretation of his position. But neither should we forget that his stated aim is one of "giving the optimist something *more* to say" (F&R: 4 (my emphasis)).

I turn now to the *Presupposition question*. What sort of power or capacity—more generally, what sort of property—do our practices of praise and blame presuppose in a so-called "responsible" agent?

Whatever sympathy Strawson has with the pessimists' camp, it evaporates on this point and begins his long campaign to wrest from them a "formal withdrawal" of their metaphysical demands. While acknowledging that morally responsible agents must have whatever it takes to genuinely *merit* praise or blame for the good or bad things they knowingly and intentionally do, Strawson adamantly denies that this can or should be spelled out in libertarian terms. As he repeatedly emphasizes, he is not even sure these terms can be given any coherent or sensible articulation.

[8] For a nice discussion of this point, see Bennett 1980: 19–20.

The crucial move involves a turn to the practical—or, as Strawson insistently says, to "the facts as we know them". His argument, which I here present in a very abbreviated and schematic form, leads to both a negative and a positive conclusion.

We begin with the commonplace that praise and blame are but exemplary instances of a whole range of "reactive" attitudes that are likewise experienced in a wide range of interpersonal encounters—so much so that it's barely conceivable to think of our human way of life without them. (Strawson's well-known list includes gratitude, resentment, hurt feelings, indignation and approbation, shame and guilt, remorse and forgiveness, certain kinds of pride, and certain kinds of love.) Yet these attitudes are distinctive—that is, marked out as a class—so far as they express two things: First, a basic sensitivity to how people are regarded and treated by one another in the context of their interactions; and, secondly, a normative demand (modulated to suit the interactions in question) that such treatment and regard reflect a basic stance of good will (F&R: 14–15). But if that's the case, it only makes sense—it's only appropriate—to direct our reactive attitudes to *agents who are capable* of understanding and living up to the normative demand expressed in these attitudes—that is, agents who are fit to be held responsible, and therefore appropriately deemed responsible, in the normative terms of our practice.[9] So far, so good—but how is this determination to be made?

It is at this point that we get a critical reassertion of the practical dimension in Strawson's thinking. For instead of launching into an *a priori* discussion of

[9] These points have been richly explored by Gary Watson (1987) and R. Jay Wallace (1994). A propos Strawson's list of reactive attitudes, it's worth noting that Wallace argues this list should be more restricted precisely because only some of these attitudes—the "central cases are resentment, indignation and guilt"—are properly viewed as essentially connected with the normative expectations we have of one another. The reason is they all involve the same propositional object—viz., the belief that x is a normatively competent agent who has violated a normative expectation/demand. Thus, in Wallace's estimation, these reactive attitudes "hang together as a class" (Wallace 1994: chs. 1 and 2). But while this is certainly true, it is unclear why reactive attitudes have to have precisely the same propositional content to hang together as a class. After all, they may hang together as a class because they share a unifying presupposition—viz., that they are only properly directed towards agents of whom it is appropriate (or "fair", as Wallace likes to say) to make normative demands. If there is good reason to see all the reactive attitudes on Strawson's list as fitting this bill, then there is good reason to count them all as reactive attitudes. I take Watson to endorse something like this view. There is another sense, too, in which these reactive attitudes may hang together as a class, despite varying in propositional content. As I explain in Section 2, having these attitudes only makes sense in the context of a normatively shaped trajectory of reactive exchange. In other words, some reactive attitudes hang together only so far as they are (seen as) normatively appropriate moves within a kind of dialogical exchange that aims, for instance, at repairing harm and/or restoring community after a wrong has been done (see too McGeer 2012).

what must be involved in having such a capacity, Strawson makes the crucial observation that we do in practice distinguish between agents who are fit to be held responsible and those who are not. Indeed, this is an essential feature of our practice—we couldn't have any such practice unless certain kinds of creatures were excluded (animals, infants, the deeply disturbed). But, admittedly, it is a feature that admits of grey areas, since it rides on a distinction that is acknowledged (in practice) to be a matter of degree.

Now why is this observation a propos? The answer is that it takes us immediately to what I call Strawson's negative conclusion.[10] The fact that we can and do make such a distinction in practice means that we're reasonably good at discerning the very property of agents that make them an appropriate target of our reactive attitudes. So this property cannot consist in the exercise of a libertarian free will, a property that—if it exists at all—certainly goes well beyond the facts as we know them.

Of course, this negative conclusion immediately invites a complementary positive conclusion—the fact that we're reasonably good at discerning this property in practice means that no metaphysical chicanery is needed to make sense of it. As philosophers, we can explore this property in more detail, and the complex constellation of features that no doubt underwrites it, simply by examining the pattern of excusing and exempting conditions that emerges in our practice.[11] For reasons I come to in the next section, I will call the property that we triangulate on in this way the property of "co-reactivity"; but whatever term we give it, this is the property that, in Strawson's view, is distinctive of responsible agency.

Finally, on the third *Metaphysical question*, which now hardly needs spelling out: Strawson simply denies the relevance of the thesis of determinism to the presupposition we make about responsible agents, as reflected in our attitudes and practices of holding responsible. The truth *or* falsity of determinism goes beyond the facts as we know them; and yet, it is the facts as we know them that we rely on to distinguish those who are fit to be held responsible from those who are not; it is therefore the facts as we know them that must bear on how we identify the property that makes for responsible agency.

The three positions I've outlined in this section are summarized in the diagram in Figure 4.1: the two classic positions that Strawson aims to

[10] This negative point is implied in much of what Strawson says, but see esp. 23–4. I defend this move in more detail in McGeer 2012.

[11] Of course, as Strawson says, the details here can—and should—be a matter of some debate. That is to say, the shape of our practices is, and should be, open to revision according to "internal" standards for modification and redirection (F&R: 23). I return to what further sense we can make of this controversial point in my conclusion.

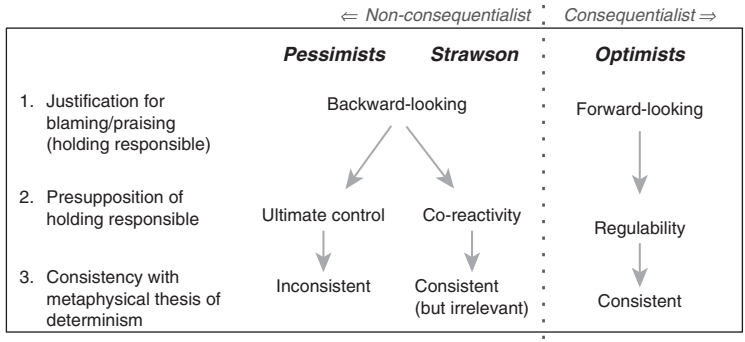

Figure 4.1 Three perspectives on holding responsible: Strawson vs. the classic views

reconcile—the optimists and the pessimists, and Strawson's purported reconciliation. I include a dashed line to indicate how these views are standardly taken to line up with respect to the issue of consequentialism. Apart from the dashed line (which in my estimation should be moved to the left), I think this representation of Strawson's view is fine so far as it goes. But it overlooks an important dimension in Strawson's picture, and I turn now in Section 2 to bringing out this dimension.

2. COMPLETING THE STRAWSONIAN PICTURE

Let me begin this part of my discussion by citing (part of) the closing paragraph to "Freedom and Resentment":

> If we sufficiently, that is *radically*, modify the view of the optimist, his view is the right one. It is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes, in regulating behaviour in ways considered desirable; or to add that when certain of our beliefs about the efficacy of some of these practices turn out to be false, then we may have good reason for dropping or modifying those practices . . . (F&R: 25 (emphasis in the original)).

I will return to this passage shortly and to the important caveat Strawson adds to these remarks to round out his view. But first I want to illustrate a number of strands in his thinking that fit with the spirit of these remarks. These strands amount to showing that he thinks of the reactive attitudes as serving a *forward-looking regulative purpose*—a purpose that appears to be ruled out in the representation of Strawson's view presented in Section 1. So my aim in this section is relatively modest: I simply want to defend

the claim that we need at least to enrich our understanding of Strawson's view. In Section 3, I will go on to consider how best to characterize this view (so understood) in the wider philosophical landscape, therein defending the more adventurous claim that Strawson is best understood as embracing a (sophisticated) form of consequentialism.

For all intents and purposes, I ended Section 1 on a somewhat promissory note. Strawson's proposed reconciliation between optimists and pessimists importantly begins with focussing our attention on what is in dispute—viz., the kind of property that agents must possess in order to make them proper or appropriate targets of our reactive attitudes and practices. Given the nature of these attitudes and practices, Strawson agrees with the pessimists that such a property must comport with our sense that reactive responses are merited or deserved; but, equally, and against the pessimists, Strawson avers, we will only respect the integrity of our practices so far as we concede that such a property is reasonably discernible in the context of our day-to-day interactions. As noted in Section 1, the property in question must consist in—or at least attest to—our *capacity* to understand and live up to the norms that make for moral community. But we have not yet seen how Strawson arrives at this conclusion, nor why we should regard this property as metaphysically undemanding—that is, as the kind of property that is not only manifested in our day-to-day dealings with one another, but also requires no exceptional metaphysical conditions to make sense of it. The reconciliation between optimists and pessimists must finally hinge on the details of this account; but we can only develop this account, in Strawson's estimation, through a careful consideration of excusing and exempting conditions. So this is the next step—and here I will simply highlight some of the main points Strawson makes in this regard.

First, a consideration of *excusing conditions* should make clear to us that our reactive attitudes are not normally or properly provoked by desirable or undesirable behaviour *per se*; rather, we take such reactions to be merited or deserved only when a person's behaviour expresses or betrays an objectionable *attitude* on their part towards other people. A person's attitude is objectionable, in Strawson's estimation, so far as that person fails to show a reasonable degree of care or concern for others (what Strawson calls "goodwill") in the context of their interactions. Thus, a person may be excused if they cause some harm accidentally or even unwittingly (though certain kinds of negligence can presumably stem from an inexcusable lack of concern); but if a person does the same thing maliciously or with an ill will, they rightly attract a reactive response from us—our resentment, indignation, and/or blame. Why is this? Strawson makes two points in this connection. The first is that it is simply a natural feature of human psychology that we care enormously about how others regard us—with "goodwill, affection

or esteem on the one hand or contempt, indifference, or malevolence on the other" (F&R: 5). But, secondly, and more importantly, this care has been elevated into a normative expectation or demand for reasonable regard; that is, the degree or kind of good will suitable (as we think) to the wide variety of relationships and interactions that make up our social world (F&R: 6). The fact that we make this normative demand of one another is reflected in the fact that we judge various reactive attitudes to be merited or deserved when the demand for good will has been flouted.

But now let us turn to Strawson's so-called *exempting conditions*. The attitude of "goodwill, its absence or opposite" is clearly not the only feature we are tracking in others, on Strawson's view, and clearly not sufficient for justifying a reactive response. In addition, we must consider why we exempt certain people from these responses—and believe we ought to exempt them—either on a temporary or a permanent basis. Again, I won't go into nuances here, but the central cases are clear. They involve people who are cognitively and affectively abnormal in various ways (perhaps they're psychotic, or deeply neurotic, or brain damaged in certain critical respects); and though these people may injure or even benefit us, we don't think they are a suitable target for our reactive responses precisely because they are "an inappropriate object of the kind of demand for goodwill or regard which is reflected in our ordinary reactive attitudes" (F&R: 7). What could make such people an "inappropriate object" of this normative demand? Strawson's thought is clear: their cognitive/affective handicap either makes them *incapable* of understanding the kind of demand expressed in our reactive attitudes or it makes them *incapable* of living up to that demand. Hence, they are unfit to be treated as "participants" in our shared moral practice—so it makes no sense to respond to them reactively. Of course, we might respond to them in all sorts of other ways: we may think it right to manage them, or restrain them, or provide them with some kind of treatment. And naturally this does not mean that they fall outside the scope of our moral regard. The point is just that we reserve our reactive responses for those whom we take to be capable of understanding and living up to the demands that we communicate through our reactive attitudes.[12]

[12] Gary Watson (1987) provides a very rich discussion of the problems and puzzles that arise from a Strawsonian account of responsibility in view of the fact that it relies so critically on the capacity to understand and live up to the normative demands we make of one another. For instance, he discusses the case of irretrievable evil, such as exemplified by the serial killer, Robert Harris, who refuses to abide by the shared normative demands of moral community. Does not this persistent refusal amount to something like an incapacity? I agree with Watson's apparently settled view that it does not—at least not obviously so. Yet Watson goes on to press the worry that Harris's refusal is understandable

So far so good. As Strawson insists, a consideration of excusing and exempting conditions makes clear to us the kind of property we take to make for responsible agency—viz., the possession of a certain sort of capacity. But now we want to know more about what it means to "possess" the capacity for understanding and living up to the demands communicated through our reactive attitudes: What feature or features must be present in an agent such that they are properly deemed to possess this capacity, whether or not they act well or badly? And why is the presence of such features in no way compromised by any lingering metaphysical worries the pessimist has tried to raise? A Strawsonian response to these questions involves, in my view, a more careful examination of what precisely we are communicating through our reactive attitudes and what we think it takes for this message to be adequately received.[13]

Begin, then, with what we are communicating through our reactive attitudes. As I've emphasized above, it's certainly part of our message that we expect—indeed, demand—that responsible agents show one another an appropriate degree of moral regard. But given that our reactive attitudes are sensitive to judgements we make about whether or not someone is a fitting recipient of these attitudes, the fact that we express them effectively communicates a good deal more. It says to their recipients that we don't despair of them as moral agents; that we don't view them "objectively"—as individuals to be manipulated or managed or somehow worked around; indeed, that we hold them accountable to a standard of moral agency because we think them capable of living up to that standard. So reactive attitudes communicate a positive message even in their most negative guise—even in the guise

in light of the horrific upbringing he suffered through no fault of his own. Does that not make us less inclined to blame him for what he does? Again, I agree with Watson that it does—but I am inclined to suggest that it makes us revisit the question of how well developed Harris's capacity for moral agency could possibly be. Nevertheless, on my account this does not argue for refusing to hold Harris responsible for what he does, since holding responsible is about working to develop an agent's moral capacities (this theme will emerge more clearly by the end of this section). Hence, Harris is exactly the sort of agent we *should* hold morally responsible, aiming thereby to develop what may be only nascent in him. (For some vindication of this perspective, see Watson's postscript to his original essay in (Watson 2004: 258–9). Thanks to an anonymous referee for pointing this out.) Still, our practices of "holding responsible", as Strawson himself notes, must be inflected to suit the moral developmental condition of the agent in question, and this may entail some partial retreat to the objective stance (as in the case of individuals who, for one reason or another, are temporarily or partially exempted from our full-dress reactive attitudes). This of course is a very large topic, and I here only indicate briefly how I would address it.

[13] I discuss these points at some length in McGeer 2012 where I introduce the notion of "co-reactivity" and trajectories of reactive exchange. Figure 4.2 is also reproduced from this paper.

of anger, resentment, and indignation. The fact that we express them says to their recipients that we see them as individuals who, *going forward,* can certainly do better in understanding and living up to the norms that make for moral community. In other words, the capacity we attribute to responsible agents, by way of our reactive attitudes, is invariably a forward-looking capacity for moral engagement and development.

Now let us turn to the other side of this communication—the recipients. Our reactive attitudes will be well targeted, I've said, if their recipients can understand this message and have a proneness—or, at least, susceptibility—to respond in ways that show normative awareness of the demands being made of them. But what will such a response involve? It may reflect some prior understanding of why their behaviour prompted the reactive attitude in question. But I don't think this is the essential thing. What's more essential is that the recipients of such attitudes understand—or can be brought to understand—*that* their behaviour has been subjected to normative review, which review now calls on them to make some normatively "fitting" response. Of course, such responses may still be many and varied. They will depend, for instance, on whether the recipient agrees with the judgement implied in the reactive attitude. For instance, in the case of anger or resentment, a recipient can show basic normative sensitivity in my sense by getting defensively indignant in return, thereby refusing (initially at any rate) to accept the moral judgement implied in the reactive attitude. However, such defensive indignation is rarely very satisfying to either party in the exchange. The reason, I suspect, is that morally capable agents have a basic human need to reach agreement on the normative significance of what they do to one another. Thus, in optimal cases, a (normatively) fitting response to anger or resentment involves parties on both sides working to understand why the original behaviour prompted a reactive response, and for the putative offender to make amends if amends are really due.

In sum, reactively responsive agents are the kind of agents that care, or can be brought to care, about living up to the demands of responsible agency that we express in and through our reactive attitudes. And by "living up to the demands", I simply mean that, however they have failed before, such agents will at least behave *reactively* in ways commensurate with treating them as responsible agents—ways that include justifying or reviewing their actions, negotiating about their meaning, and (in cases of genuine offence) coming to terms with what they might owe others by way of contrition, apology, and commitment to reform. Hence, the kind of responsiveness we look for in responsible agents—that is, agents who we take to be appropriate targets of the reactive attitudes—can now be summed up in a single word: *co-reactivity*. They are co-reactive agents—"co", because they show, by virtue of their *own* reactive responses (some better, some worse),

a basic normative sensitivity to the reactive attitudes of others, and hence a susceptibility to be engaged in the kind of normative exchanges that enlarge an agent's moral understanding. Co-reactivity is thus, on my reading of Strawson, the bedrock feature of responsible agency: it is a feature that is implicated in the attribution we make to others of a capacity to understand and live up to certain normative demands when we make them a target of our reactive attitudes; and it is a feature that is in no way threatened by metaphysical speculations one way or the other.

Having identified this property, we can now highlight two further features of reactive attitudes that make sense of the critical role they play in our interpersonal lives. These features are implicit in Strawson's discussion, but I don't think they get nearly enough attention.

First, there is a tendency, no doubt encouraged by the name Strawson gave them, to focus on the fact that reactive attitudes are backward-looking responses to the actions and attitudes of others. But, as attitudes themselves, they naturally prompt reactive responses in turn. After all, as Strawson points out, our reactive responses reflect the fact that we care enormously about what attitudes others manifest towards us, and this will be true—perhaps even more true—when the attitudes in question are themselves reactive attitudes: attitudes that, in their nature, are commenting—with approval or disapproval—on the quality of our moral agency. So reactive attitudes are backward-looking responses to the actions and attitudes of others, to be sure. But, more importantly, they have a forward-looking dimension, serving—and, indeed, *aiming*—to elicit some further reactive response from the individuals to whom they're directed. This explains why particular reactive attitudes tend to persist—and we judge it appropriate for them to persist—until a suitable response is forthcoming. Moreover, it is this forward-looking dimension that explains how reactive attitudes can play a critical role in scaffolding—that is, developing and supporting—our capacities as moral agents: namely, by prompting reactive responses in others that are (judged to be) normatively appropriate.

This leads to a second important observation: I have said that reactive attitudes will function successfully in their scaffolding so far as they prompt normatively appropriate responses from others. But since this is part of their aim—to *elicit* such responses—when that aim is accomplished, these reactive attitudes are normatively answered and suitably transformed, replaced by new reactive attitudes, that are themselves appropriate responses to the reactive responses prompted by the original reactive attitudes. In other words, reactive attitudes perform their scaffolding role so far as they are normally embedded in normatively meaningful trajectories of reactive exchange. These trajectories are actually what give the reactive attitudes that constitute them the meaning and power they have. Forgiveness is a good

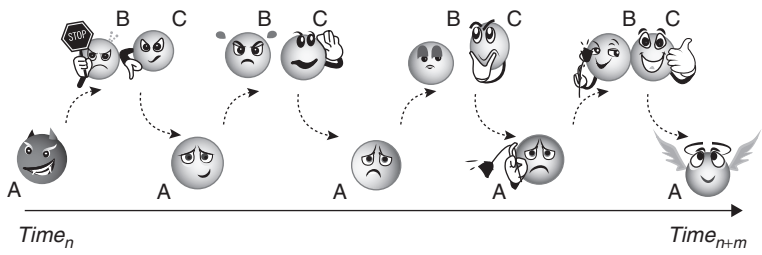

Figure 4.2 A sample trajectory of reactive exchange: *The forgiveness trajectory*

example (see Figure 4.2).[14] Forgiveness is a reactive attitude that serves, among other things, to reaffirm the moral competence of the individual to whom it is directed. But it only makes sense as a reactive attitude—and only has the power it does—so far as it normally comes at the end of a trajectory of reactive exchanges occurring principally between a victim and a wrongdoer, but often involving the reactive responses of bystanders as well. Hence, if we want to understand how reactive attitudes play a constructive role in making and sustaining moral community, we need to understand the normative trajectories of reactive exchange in which their "call and response structure" finds a natural home.[15]

Now how do these observations concerning both the backward- and forward-looking dimensions of reactive attitudes tie into the quote with which I began this section? Let me return to this passage now in full:

> If we sufficiently, that is *radically*, modify the view of the optimist, his view is the right one. It is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes, in regulating behaviour in ways considered desirable; or to add that when certain of our beliefs about the efficacy of some of these practices turn out to be false, then we may have good reason for dropping or modifying those practices. What *is* wrong is to forget that these practices, and

[14] I use the example of forgiveness to illustrate the trajectory-dependent qualities of reactive attitudes for two reasons: (1) Such trajectory-dependence comports with Strawson's own rather brief observations concerning the conversational reactive dynamic in which the asking for, and offering of, forgiveness is situated: "To ask to be forgiven is in part to acknowledge that the attitude displayed in our actions was such as might properly be resented and in part to repudiate that attitude for the future (or at least for the immediate future): and to forgive is to accept the repudiation and to forswear the resentment" (F&R: 6); and (2) such trajectory-dependence also partly explains why Strawson would—and I think should—have included forgiveness on his list of reactive attitudes (*pace* Wallace—see n. 9).

[15] I borrow the term "call and response structure" from Coleen Macnamara (Macnamara preprint 2013).

their reception, the reactions to them, really *are* expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. Our practices do not merely exploit our natures, they express them. Indeed, the very understanding of the kind of efficacy these expressions of our attitudes have turns on our remembering this. When we do remember this, and modify the optimist's position accordingly, we simultaneously correct its conceptual deficiencies and ward off the dangers it seems to entail, without recourse to the obscure and panicky metaphysics of libertarianism (F&R: 25, underlining is my emphasis, italics are in the original).

As I read this passage, it is clear that Strawson views the reactive attitudes and practices as having a regulative rationale—otherwise, we should drop or modify them. But his point is to emphasize that the *mechanism* of regulation makes all the difference in the world to explaining both the *kind* of regulative efficacy these practices have and the *normative weight* we commonly invest in them. These points are not unrelated. The normative weight, as I understand it, signals our commitment to treating others as autonomous agents, capable of understanding and living up to the normative demands we place on one another in and through our reactive practices. And the kind of efficacy these practices have stems from the fact that we naturally care about one another's attitudes; indeed, we care enough to persistently engage with one another in the kind of exchanges that (implicitly or explicitly) target the need for attitudinal change in addition to behavioural reform. In short, the kind of regulation at issue is regulation by way of shaping and developing one another's capacities to act in ways that are commensurate with the normative demands we make of one another.

In sum, we are now in a position to see how Strawson's view constitutes a *genuine* reconciliation of the views of the optimist and the pessimist—one that takes something critical from each side. From the optimist, he takes what we might call the ultimate or external rationale for our reactive attitudes and practices—namely, the regulation of behaviour; and from the pessimist, he takes what we might call the proximate or internal justification for targeting one another with such attitudes—namely, a presumed capacity as responsive and responsible agents to be suitably moved and motivated by the normative demands such attitudes express.

3. STRAWSON'S CONSEQUENTIALISM

My goal in Section 1 was to present a characterization of Strawson's view that I take to be fairly standard and uncontroversial. In Section 2, I emphasized certain strands in Strawson's thought that underscore not just the communicative, but also the forward-looking or "scaffolding" dimensions of our

reactive attitudes and practices. Now, in this final section, I want to return to the question of how best to characterize Strawson's position in the wider philosophical landscape; and my argument will be that we make best sense of his position—indeed, show it to be less vulnerable to persistent objections at certain crucial points—if we take him to be defending an account of responsibility within a broadly consequentialist framework.

As I said in my introductory remarks, I'm aware that this interpretation runs counter to an impressive tradition of Strawsonian scholarship—one that has certainly influenced and deepened my own understanding of Strawson's views, and to which I am greatly indebted. Nevertheless, I see this scholarship as going unnecessarily overboard in resisting, on Strawson's behalf, any unhappy taint (as it's often perceived to be) of consequentialist thinking. But let me point out, referring again to the text, that Strawson takes himself to be castigating the optimists merely for what he calls their "characteristically *incomplete* empiricism, a *one-eyed* utilitarianism" (F&R: 23 (my emphasis)). This suggests that it is not the utilitarianism *per se* that he finds objectionable, but rather the optimists' persistent insensitivity to *all* "the facts as we know them".[16] Recovery of these facts paves the way for a "radically modified" version of the optimists' position, to be sure; but in my estimation, it paves the way for a particularly sophisticated form of consequentialism, which some authors defend today under the rubric of "indirect" or "restrictive" consequentialism (Railton 1984; Pettit and Brennan 1986; Pettit 2012).[17] This is the position I take Strawson to be embracing.

Before characterizing what I take this form of consequentialism to be, let me just say a few words about the basic framework within which Strawson's view must be situated if this interpretation is to be made good. The central tenet of (act) consequentialism, as I understand it, is simply this: what makes a choice right—and therefore normatively justified—in any field of action ultimately comes down to whether it promotes the general (or agent-neutral) good as well as, or better than, any alternative. The good

[16] The terms "utilitarianism" and "consequentialism" were not well distinguished when Strawson wrote "Freedom and Resentment", so I take his concern to be, more generally, with a certain form of consequentialist thinking.

[17] I use this term with some caution, as it seems to mean different things to different philosophers. In particular, some take indirect consequentialism to refer to some version of "rule consequentialism", where the rightness of an act is not determined directly by its consequences, but only indirectly by, say, conformity to a rule that in general produces the best consequences (for discussion, see Sinnott-Armstrong 2014). This is *not* how I shall be using the term. I define it more precisely later, but here I am following the lead of Railton and Pettit where the indirection refers to how agents should (generally) deliberate about options, as opposed to the criterion that determines which action is right.

may be variously defined, of course, whether as utility or in some other way. Indeed, there may be plural sources of value, leading to incommensurable goods, and therefore—at least in some domains—choices between alternatives that agents must regard as indeterminate. And finally, as even the critics of consequentialism make clear, the good need not be extrinsic, or causally detached, from the means whereby it is produced; the good features of an act—for example, the kindness of a compassionate act, or the restfulness of the act of lying down—may count among its (property-manifesting) consequences.[18] (I will have cause to return to this point later.)

Now for the crucial distinction I have in mind between "direct" and "indirect" (or "restrictive") consequentialism. Writing in 1874, Henry Sidgwick made the following wise observation:

> It is not necessary that the end which gives us the criterion of rightness should always be the end at which we consciously aim: and if experience shows that the general happiness will be more satisfactorily attained if men frequently act from other motives than pure universal philanthropy, it is obvious that these other motives are reasonably to be preferred on utilitarian principles (Sidgwick 1966: 413).

Here Sidgwick suggests that two-eyed utilitarians may have to countenance a divergence between the *motives* for which people act and even the principles that ground their deliberation, on the one hand, and the goal that *justifies* their acts in order for that goal to be optimally secured, on the other. This potential divergence is what makes for the difference between direct and indirect consequentialism, as I will understand these terms here.

Direct consequentialists argue that, in deciding how to act, agents need only think about the good they can do and let this principle guide their decisions directly. In other words, their only motive, their only principle in deliberation, may be to produce as much good as possible in a choice. By contrast, *indirect* consequentialists maintain that in many cases this would be counterproductive because certain goods can only be reliably produced—indeed, perhaps only produced at all (and this represents a departure from Sidgwick's view)—if agents are *not* directly guided by the aim of producing such goods. The hedonistic paradox provides a nice example;[19] but perhaps the best examples are with certain practice-dependent goods.

[18] See, for instance, Williams' contribution in (Smart, J. J. C. and Williams 1973). See too (Anscombe 1958). Thanks to Bennett Helm and Philip Pettit for helping me to see the importance of this point.

[19] The hedonistic paradox suggests that we can only achieve happiness when we don't aim at it directly: happiness is "essentially a by-product" of other activities that we engage in. The language of essential by-products comes from (Elster 1983), who provides a rich discussion of this phenomenon.

Consider, for instance, the practice of friendship.[20] Most consequentialists will recognize that, in general terms, the world is a better place so far as there are people who form friendships and abide by the norms that the practice of friendship invariably entails. A world without friendship would be a world that is considerably impoverished from our human point of view. But to abide by the practice of friendship is to feel, think, and deliberate like a friend. Among other things, this means giving a certain priority to the needs and interests of one's friends, and responding to those needs and interests, more or less spontaneously, out of one's particular care, concern, and affection for them. Manifestly, this rules out a more calculative deliberation that considers the overall neutral good that one's response to a particular friend would serve, given the relative importance of friendship in general, and of this friendship in particular. Such calculative deliberation involves, in the cutting terms of Bernard Williams, one thought (indeed, many thoughts!) too many.[21] Thus, a true consequentialist will have reason to avoid thinking like a consequentialist in certain relationships and situations—for example, within the practice of friendship; in these matters, the consequentialist will have reason just to think like a friend, for this is the only way to produce the good that genuine friendship brings into the world.

But, now, how can a consequentialist really take the demands imposed by this kind of indirection on board? Isn't there a self-defeating dilemma lurking in the wings that undermines the stability of any such position? Consider, again, the practice of friendship:

> On one horn of the dilemma, the indirect consequentialist locks herself into the practice, as it were—into thinking like a friend. But, now, in putting such a premium on thinking like a friend, she may end up doing things that betray the general good that friendship supposedly serves. She may allow her love or loyalty to lead her into activities that are in no way defensible in consequentialist terms.[22] On the other horn of the dilemma, the consequentialist monitors what she is doing whenever she acts as a friend to ensure the good-making features of her activities are not so betrayed. But in this case, it seems she must engage once again in the kind of deliberation that is inimical to responding as a friend. So she fails to produce the good in question, and so fails to adhere to what her indirect consequentialism sensibly dictates.

How is this dilemma to be avoided? If consequentialists are not to become enslaved to the practice as such, they need to be able to identify cases where

[20] This example comes from Pettit 2012.

[21] For Williams' discussion of "one thought too many" and its relation to choice and action, see (Williams 1981: ch. 1).

[22] This is the danger implicit in "rule consequentialism"—and for this reason, a kind of position that indirect consequentialists argue we should avoid (Pettit and Brennan 1986; Pettit 2012).

the good is better served by exiting the practice. But if they are not to undermine the good that is served by participating in the practice, they cannot continually monitor the likely upshot of their choices, as from a calculating consequentialist perspective.

Philip Pettit recommends one compelling strategy by which indirect consequentialists can meet these constraints (Pettit 2012). It involves relying (if possible) on circumstances alerting individuals to cases where exit may be the proper option; to rely on there being external cues—a red warning light, as it were—that prompt rethinking in such cases. In other words, indirect consequentialists must rely on the possibility of "outsourcing control"—letting cues from the environment (or "red lights") alert them to when the situation does not call for business as usual. In a memorable example from Dean Cocking and Jeanette Kennett (Cocking and Kennett 2000): If a friend asks you to move an apartment, you will generally want to help—without any soul-searching second thoughts. But if a friend asks you to move a body, the red lights will certainly go on—things are not as usual; and they surely ought to give you pause.[23]

To sum up the points I have made thus far: (i) Consequentialists argue that the right option in any choice is that which suitably promotes a relevant neutral good. (ii) In many cases, as in the practice of friendship, the promotion of the relevant good requires not focussing on achieving this good directly, but rather simply operating in accord with appropriate practice-dependent norms. (iii) But participating in the practice can still be subjected to consequentialist control by reliance on the red lights that circumstances can generate, alerting us to the fact that, in a particular case, it may be best to exit the practice. With these points in hand, we can return to Strawson's analysis of the phenomenon of holding responsible, as this is embodied in our complex and norm-governed web of reactive attitudes and practices.

With respect to (i), the attitudes and practices of holding responsible—that is to say, our reactive attitudes and practices—produce a clear agent-neutral good by Strawson's own insistence: the good of "regulating behaviour in ways considered desirable". Moreover, this good is not simply a welcomed by-product of attitudes and practices that should be maintained for other reasons. For, as Strawson also insists, if they turn out *not* to be efficacious in producing this good, then we have "good reason for dropping or modifying these practices" (F&R: 24; see also 22).

[23] The expressions "red lights" and "outsourcing control" come from Philip Pettit, who also cites the example from Cocking and Kennett (Pettit 2012). For further discussion and elaboration of these ideas, see Pettit 2015 (forthcoming).

But, now with respect to (ii), the promotion of this good would be jeopardized if our treatment of others (and, indeed, ourselves) became overly dominated by this regulative concern. In particular, it might encourage a retreat from regarding one another as responsible agents, as individuals who are capable of understanding and living up to the demands that are expressed in and through our reactive exchanges. And this in turn would compromise the goal of regulating behaviour *in ways considered desirable*—that is, in ways, as I have just emphasized, that work to develop people's moral capacities, including their capacity to engage in appropriate norm-governed behaviour.

This is precisely the trap that optimists fall into. They replace any substantive notion of a morally responsible agent with an agent—an object—that can be appropriately conditioned by the crude application of carrots and sticks. In their vision, reactive attitudes become mere tools—"devices we calculatingly employ for regulative purposes" (F&R: 25); and, by that token, must surely come to be seen as dispensable if other tools should prove more effective in achieving the end of social regulation. But such a vision not only belies our humanity, suggesting that we could give up on attitudes and practices that are an essential feature of interpersonal relationships as we know them; it also ignores the vital fact that we human beings are, by nature and nurture, the sort of creatures that are deeply responsive in thought and action to the moral demands we place on one another; indeed, we are the sort of creatures that are especially responsive when such demands are couched in a way that expresses an acknowledgement of, and respect for, our capacities to reason and respond to one another as responsible agents. This is the vital thing that is embodied in our reactive attitudes and practices, the vital thing that we cannot lose sight of without losing, in one blow, something that is essential to our humanity and that provides a critical resource for regulating behaviour in the indirect way of scaffolding one another's moral capacities. Hence Strawson's rebuke to the optimists: they put their concern for achieving the good of social regulation ahead, or indeed in place, of a proper concern with how that regulation is achieved—in and through the complex web of our reactive exchanges; hence, they misunderstand the nature of the good at which they supposedly aim.

Yet, now finally with respect to (iii), how can we fine-tune our reactive attitudes and practices such that we don't become slavishly locked into them—always treating one another as responsible agents, whether or not such treatment yields the consequentialist good of regulating behaviour in ways considered desirable? How are we able to track this consequentialist good in our practice without making behavioural regulation our primary focus and concern? Here Strawson relies on our sensitivity to various

exempting conditions. When someone is appropriately exempted from the practice (they are mad or delusional or cognitively incapacitated in some way that makes them incapable of understanding and/or living up to our normative demands), this fact will be more or less salient in their demeanour and in their responses to us. The "red lights" will generally go on in appropriate circumstances, and we will retreat to something like the objective stance, considering how best to interact with them as subjects of "treatment", who are not wholly fit to participate in full-dress reactive exchanges.[24]

Let me complete this characterization of Strawson's view as a "red lights" (or "outsourcing") form of indirect consequentialism by adding a proviso. This characterization is in danger of being misleading in one important respect. It may suggest that those who participate in a practice-like friendship, or in the co-reactive practice that Strawson has in mind, have got to embrace something like a split personality, forswearing any reflection on the good that the practice serves—short, at least, of the red lights going on—in order not to compromise that very good itself. But this suggestion is misleading. Consistently with the story told, participants in the relevant practice can avow the good that is the goal of the practice so long as they understand this good in such a way that the appropriate means is built into a specification of the good itself—that is, the means are a non-detachable part of the good to be produced.[25] Thus, the relevant good is not just regulation-by-any-means, or helping a friend out-of-any-motive, but rather regulation-in-the-manner-of-reactive-engagement, or helping-a-friend-out-of-friendship. The point is not that participants cannot be mindful of the goal served in the practice, but rather that they cannot focus exclusively on that goal, seeing it as something distinct from the practice that might be served in any of a number of ways, including ways that offend against the norms of the practice itself.[26]

[24] This is not to say the "red lights" will invariably go off. There may be controversial cases where we don't know what to think; or cases, such as psychopathy, where it might go very much against the intuitive grain to judge the individuals in question as non-responsible. But, as in all things, we can become better informed about psychiatric conditions, and how these might or might not be compromising for moral agency. And as we become better informed, the norms of our reactive practice, including taken-for-granted exempting conditions, are likely to change, and with them our capacity to recognize those conditions *in situ*. I touch on the issue of how our reactive practice might develop and change again in my concluding remarks.

[25] As I emphasized at the outset of Section 3, even critics of consequentialism allow that some goods can be like this. See also Moore on the idea of an "organic good" (Moore 1960: ch. 3).

[26] A further example may make this point vivid (this example comes from McGeer 2013). In engaging with others as rational deliberative agents, our goal is often to change

Taking these points together, it seems clear that Strawson embraces a form of indirect consequentialism such as described in this section. He thereby joins a long tradition of consequentialist thought, while being original in the emphasis he gives to the need to abide by the internal norms of the practice he celebrates.

4. CONCLUDING REMARKS

But now why place so much emphasis on the consequentialist aspects of Strawson's thought? What value does this add to the rich array of scholarship and commentary already generated by "Freedom and Resentment"? As I said at the outset, it seems to me to provide a needed corrective to a dominant strand of this scholarship; but, more significantly, it adds some depth and resilience to certain points that Strawson makes—points that critics have suggested lack appropriate grounding or foundation. Here very briefly, and by way of conclusion, are two of the points I have in mind.

The first criticism arises at a practical level and runs something like this. While Strawson emphasizes myriad *internal* standards according to which particular manifestations of our reactive attitudes and practices can be subjected to review (the excusing and exempting conditions), there are no *external* standards that can be brought to bear on the overall shape of our attitudes and practices themselves—allowing us to ask, for instance, whether the particular excusing and exempting conditions we accept are fair or adequate; or whether our local reactive practices should in some ways be reformed. After all, it seems that for Strawson, our reactive attitudes and practices are simply a given that we must accept as part of our human, and perhaps in some respects culturally specific, form of life.

Now this criticism surely misses an important aspect of what Strawson has to say. Certainly in the final pages of "Freedom and Resentment", he

their minds—to ensure that others believe something that we regard as true, rather than something that we regard as false. Moreover, there is nothing objectionable about keeping this goal in mind when we offer them arguments and evidence that we hope will persuade them to abandon their benighted beliefs. Keeping this goal in mind only becomes objectionable when we take the goal itself to provide us with a certain kind of entitlement—the entitlement to adopt any means available to make them change their minds: e.g., a blow to the head, a hypnotic drug, or some other form of neural tinkering. For these means are simply inimical to the practice of engaging with others as rational deliberative agents. The take-away lesson is simply this: indirect consequentialists need not insist that participants in a practice keep the good they seek to attain through participating in that practice fully out of view, so long as they regard the good they seek to attain as essentially practice-dependent.

makes clear that our reactive attitudes and practices are not immutable; that they can—and should be—subject to "modification and redirection". But this reformist stance makes no sense at all without embracing a standard by which such attitudes and practices could be comparatively assessed. Yet from where is this standard to come? Once again, Strawson's consequentialism comes to the rescue. Reactive attitudes and practices can be comparatively assessed in light of their aptness for producing a certain good; and it seems clear from what has gone before that the good in question is the good of regulating behaviour *by means of developing and/or supporting people's moral understanding*. Interpreting Strawson in this light not only saves his view from a certain naturalistic complacency, it provides a generative research programme, for instance in the field of criminal justice, where there is much debate concerning the appropriate institutional expression of our reactive attitudes and practices.[27]

Here is a second, more philosophical worry. There is a standing complaint that Strawson does not adequately address the foundational concern raised by the spectre of determinism for the *entire fabric* of our reactive attitudes and practices—a concern that suggests we should simply abjure these practices *holus-bolus* insofar as they presuppose too demanding a conception of human agency and responsibility. Now the entire point of "Freedom and Resentment" has been to argue that our ordinary conception of agency and responsibility, as embodied in these attitudes and practices, is not metaphysically demanding—that is, it is not demanding in such a way as to be threatened by the thesis of determinism. But suppose the sceptic remains unconvinced, seeing ordinary excuses and exemptions that are part and parcel of our practice as somehow making superficial distinctions among the ways in which agents operate that really don't count for much, morally speaking, if determinism is true. In that case, shouldn't we abandon all of our reactive attitudes and practices?

Strawson's response to this challenge has been exhaustively discussed and generally found wanting. Viewing it as a demand for the "rational justification" of our reactive attitudes and practices, he gives what amount to a two-part reply (F&R: 13). The first part insists that our "natural human commitment" to these attitudes and practices is so deep and thorough-going, so much a part of the fabric of human life, that it is simply not open, in this foundational sense, to rational review. Hence, the call to rationally justify these attitudes and practices is simply otiose. Not surprisingly, perhaps, many critics regard this part of Strawson's response

[27] For more on putting Strawsonian ideas to work in the context of criminal justice, see McGeer 2012.

as philosophically dissatisfying: it simply sidesteps the deep question that is at issue.

What about the second part of Strawson's response? While this seems to be a bit more weighty, philosophically speaking, it has something of the flavour of a bait and switch. Strawson writes: "if we could imagine what we cannot have, viz. a choice in this matter, then we could choose rationally only in the light of an assessment of the gains and loses to human life; its enrichment or impoverishment; and the truth or falsity of a general thesis of determinism would not bear on the rationality of this choice" (F&R: 13). Why the flavour of a bait and switch? The thought seems to be that Strawson is simply skirting the substantive normative issue: The pessimist wants to know whether it's *fair* or *normatively* appropriate—in *that* sense justified—to blame people for their wrongdoing; and a calculus of gains and losses to human life seems quite irrelevant to this concern.

Of course, at this stage in the dialectic, Strawson has already addressed the fairness question by showing where and how it substantively arises—viz., in the context of our reactive exchanges, where various excuses and exemptions are appropriately considered, and where the metaphysical thesis of determinism is simply beside the point. Even so, one might wonder, has the pessimist's question about fairness or normative justification really been fully addressed? Couldn't the pessimist simply raise the issue at a higher level of abstraction, so to speak—that is, with respect to our reactive attitudes and practices *as a whole*? Is it fair or morally appropriate that reactive attitudes and practices are part and parcel of our human way of life (despite what is, perhaps, their inevitability)? This question does not seem to be a demand for the *rational* justification of our attitudes and practices; it's rather a demand for their *normative* justification. And Strawson's rather cavalier response not only appears off topic, it further smacks of a kind of moral complacency.

However, once we interpret Strawson as working within a broadly consequentialist framework, this impression of irrelevance and/or moral complacency simply goes away. After all, within this framework all normative questions are to be addressed by considering (in a rational way) how the acts or practices in question measure up against putative alternatives in terms of their relative production of the good, however that is specified. We have already seen how this played out with regard to the internal justification (as we might call it) of particular manifestations of our reactive attitudes and practices: on my reading of Strawson, they are justified just to the extent that they promote the good of regulating behaviour by way of scaffolding the kind of moral agency that is critical to the warp and weave of a recognizable human society. But the pessimist seemingly wants more: the pessimist

wants to know what justifies these attitudes and practices *as a whole*; what justifies this reactively permeated human form of life (*modulo* the truth of determinism)?

To take this question seriously, the consequentialist must again have recourse to a consideration of the putative good that is thereby promoted, relative to engaging in some alternative form of life. But what kind of good could we be talking about at such a fundamental level of consideration? As I read it, the answer implicit in Strawson's response is that it is the kind of good that ought to be self-evident when we contemplate a social life replete with our human form of relationships and commitments, as against one that is stripped of all of that. *Au fond*, it is the kind of good that flows from living our human kind of life according to our nature as normatively responsive creatures. Call it "human flourishing" for want of a better term. Thus, when the pessimist presses her justificatory demand, there seems to be little left to say. For when we translate this demand into a question the consequentialist can make sense of, it has a strangely disconnected or other-worldly ring to it: "Is the apparent 'good' of being involved in the rich variety of interactions and relationships that scaffold our moral agency and make for moral community *really* a human good to be promoted?" Such a thin-sounding question calls for a thin-sounding answer, and that is precisely what Strawson supplies. Only now I think we're in a position to see the force of it (and here I am more or less paraphrasing the passage I referred to above): "If a positive answer to your question is not simply obvious in light of our 'natural human commitment' to the practices in question, then the only way to address it 'rationally' is via 'an assessment of the gains and loses to human life; its enrichment or impoverishment . . .' " (F&R: 13).[28]

To my ear, this final and telling invocation of the consequentialist criterion (even in what Strawson considers the outer reaches of sensible philosophical discourse) makes pretty clear where his normative allegiances lie. And, in light of this, I don't think he can be charged with the kind of theoretical or normative inadequacy that many have seen in his work—though some philosophers may be tempted to think that the "taint" of consequentialism is hardly worth the price. Needless to say, I am not of their number.

[28] To which he then rightly adds: "the truth or falsity of a general thesis of determinism would not bear on the rationality of this choice" (F&R: 13).

References

Anscombe, G. E. M. (1958). "Modern Moral Philosophy." *Philosophy* 33(124): 1–19.

Bennett, J. (1980). Accountability. *Philosophical Subjects: Essays presented to P.F. Strawson*. Z. V. Straaten. Oxford, Clarendon Press: 14-47.

Cocking, D. and J. Kennett (2000). "Friendship and Moral Danger." *The Journal of Philosophy* 97(5): 278–96.

Darwall, S. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. (Cambridge, MA: Harvard University Press).

Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. (Cambridge: MIT Press).

Elster, J. (1983). *Sour Grapes*. (Cambridge: Cambridge University Press).

Hart, H. L. A. (1985). Legal responsibility and excuses. *Punishment and Responsibility: Essays in the philosophy of law*. Oxford, UK, Oxford University Press: 28-53.

Macnamara, C. (preprint 2013). "'Screw You!', & 'Thank you'." *Philosophical Studies* 165(3): 1–22.

McGeer, V. (2012). "Co-reactive Attitudes and the Making of Moral Community," in R. Langdon and C. Mackenzie, eds, *Emotions, Imagination and Moral Reasoning*. (New York: Psychology Press), 299–326.

McGeer, V. (2013). "Civilizing Blame," in J. D. Coates and N. A. Tognazzini, eds, *Blame: Its Nature and Norms*. (Oxford: Oxford University Press), 162–88.

Moore, G. E. (1960). *Principia Ethica (1903)*. (Cambridge: Cambridge University Press).

Nowell-Smith, P. (1948). "Freewill and Moral Responsibility." *Mind* 57(225): 45–61.

Pettit, P. (2012). "The Inescapability of Consequentialism," in U. Heuer and G. Lang, eds, *Luck, Value and Commitment: Themes from the Ethics of Bernard Williams.* (Oxford: Oxford University Press) 41–70.

Pettit, P. and G. Brennan (1986). "Restrictive Consequentialism." *Australasian Journal of Philosophy* 64: 438–55.

Railton, P. (1984). "Alienation, Consequentialism, and the Demands of Morality." *Philosophy & Public Affairs* 13(2): 134–71.

Scanlon, T. M. (1988). "The Significance of Choice," *The Tanner Lectures on Human Values* 8: 149–216.

Schlick, M. (1939). *The Problem of Ethics*. (New York: Prentice-Hall).

Sidgwick, H. (1966). *The Method of Ethics*. (New York: Dover).

Sinnott-Armstrong, W. (2009). "Consequentialism." *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/spr2014/entries/consequentialism/>.

Smart, J. J. C. (1961). "Free-Will, Praise and Blame." *Mind* 70(279): 291–306.

Smart, J. J. C. and B. Williams (1973). *Utilitarianism; For and Against* (Cambridge: Cambridge University Press).

Strawson, G. (1994). "The Impossibility of Moral Responsibility." *Philosophical Studies* 75(1): 5–24.

Strawson, P. F. (1961). "Social Morality and Individual Ideal." *Philosophy* 36(136): 1–17.

Strawson, P. F. (1974). "Freedom and Resentment," in *Freedom and Resentment and Other Essays*. (London: Methuen), 1–25.

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

Watson, G. (1987). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme," in F. Schoeman, ed., *Responsibility, Character and the Emotions: New Essays in Moral Psychology*. (Cambridge: Cambridge University Press), 256–86.

Watson, G. (2004). *Agency and Answerability: Selected Essays*. (Oxford: Clarendon Press).

Williams, B. (1981). *Moral Luck*. (Cambridge: Cambridge University Press).