Freedom and Forgiveness

1
Introduction

"Freedom and Resentment" is a paper I return to again and again. I think it's a really fascinating, deep, subtle, incredibly important[1] and sometimes really quite annoying paper. Sometimes I return to it because I'm fascinated by its depth and subtlety, and sometimes because I just can't work out what the argument is supposed to be.[2] "Freedom and Resentment" relates, more or less directly, to two central topics in my research, which I have not previously brought into connection with each other, and between which a connection may seem unlikely: forgiveness and Kant's transcendental idealism. The paper relates to forgiveness, because Strawson introduces the now famous idea of 'reactive attitudes,' which include attitudes like gratitude and resentment; this idea is plays an important and influential role in the literature on forgiveness. Kant's transcendental idealism relates to "Freedom and Resentment" because Strawson's paper is an attempt to dissolve (or to side-step) the problem of free will, and there are some ways in which his strategy for doing this has been seen as similar to Kant's attempt to dissolve (or side-step) the problem in the first *Critique*, which depends on his transcendental idealism. Strawson aims to bring about a reconciliation between opposing sides in the free will debate through appealing to two different points of view we can adopt on the world: the 'objective' view, and the 'participant' view. This contrast has seemed to many to be similar to Kant's contrast between two ways of considering the world (as his transcendental idealism is often understood), which is central to his resolution of the free will problem.[3] In this paper I pursue Strawson's project in "Freedom and Resentment" by developing the connection between forgiveness and free will. I argue that Strawson's notion of reactive attitudes is helpful for understanding forgiveness, and that thinking about forgiveness has implications for how we understand the kind of free will reactive attitudes see us as having. I argue that this does not easily fit with the common Humean naturalist compatibilist reading of Strawson's strategy, and I suggest an alternative, broadly Kantian interpretation.

In the rest of this section I sketch Strawson's strategy in the paper, and note some different ways in which it can be interpreted. In particular, I note both Humean and Kantian strands in the argument. In section 2, I argue that Strawson's notion of reactive attitudes, with their complex intentional content, is helpful for making sense

---

1 As Paul Snowdon says in his biographical note about Strawson, it is simply staggering to think about how influential the paper has been has been, not just in the free will debate, but also in moral philosophy and philosophy of emotion, given that it is one of only two papers Strawson wrote touching on moral philosophy, and was in no way part of his central philosophical concerns. (Snowdon, 240) Strawson himself says of "Freedom and Resentment' and 'Social Morality and Individual Ideal' (published one year later): 'Between them, these two papers effectively embody all I have thought or have to say in a philosophical area which, important as I recognize it to be, I have never found as intellectually gripping as those to which I have given more attention" (Strawson 1998: 11). Similarly, many of his philosophical interlocutors do not see it as central to his work. Clifford Brown's book 2006 *Peter Strawson* does not mention "Freedom and Resentment," focusing on Strawson's work on philosophy of language, philosophical logic and metaphysics.

2 Strawson once told me that he wrote it in one draft.

3 For example, David Pears says of Strawson that he argues "in the Kantian tradition, for a double-aspect theory—we can see an agent objectively, as a cog, or we can see him personally, as an accountable originator" (Pears 1998:247.)

of forgiveness. A central difficulty with the notion of forgiveness is the idea that the resentment which it overcomes is warranted. Many philosophers have thought that this creates at least *prima facie* difficulties for the rational and moral acceptability of the shift in affective appraisal of the wrongdoer that forgiveness involves. I argue that understanding resentment as essentially affective is crucial for making sense of this: it enables us to see the resentment forgiveness overcomes as warranted but not obligatory. However, I argue, in section 3, that this is also not sufficient to solve the problem, and more work is needed to explain the content and justifiability of the affective shift which forgiveness involves. In keeping with Strawson's idea that reactive attitudes make sense only from the participant view, I argue that we cannot make sense of the shift in affective appraisal of the wrongdoer that forgiveness involves while understanding psychological states from the objective view, as causes of action. I suggest that reactive attitudes contain a particular view of the nature of intentional agency which does not see reasons as causes. This creates difficulties in understanding how such explanations involve responsibility. To make a start in explaining this, in section 4, I turn to Kant, and sketch the role of transcendental idealism in Kant's account of the free will problem in the first *Critique*. I present a way of seeing Kant's account according to which it holds that freedom requires a different kind of causal explanation to that given by science (Strawson's 'objective' view), and the legitimacy of which is based on our view of others as responsible agents, not on a non-determinist metaphysics. In section 5 I suggest that this view fits with, and makes sense of, Strawson's argument that the objective and participant attitudes don't rule each other out, and the way we are committed to thinking of persons in order to make sense of forgiveness. In section Section 6, I argue that this helps us to make sense of forgiveness.

Strawson's aim in the paper is to effect some sort of reconciliation between those who argue that if determinism were true, it would rule out freedom (represented by a character Strawson calls the 'pessimist'), and the denial of this (the view of the 'optimist'). His strategy is based on a contrast between two different views we can take on the world, which he calls the 'participant' and the 'objective' views, which involve different kinds of explanations of the actions of agents.[4] From the objective view, we see other people with the detached attitude of science, as pieces of mere nature, parts of causal chains. We see people, Strawson says, as objects of social policy, or pieces of the world to be causally manipulated: "as a subject for what, in a wide range of sense, might be called treatment…to be managed or handled or cured or trained."[5] The participant attitude, in contrast, is not detached; it is situated in the context of relationships. From the participant attitude we see the actions of other persons as making them appropriate objects of reactive attitudes like resentment and gratitude, which are responses to the attitudes of good or bad will towards us exhibited in another's actions. The content of these attitudes, Strawson thinks, cannot be captured from the objective view. The idea of moral responsibility—the responsibility for which we need the idea of free will—is bound up with seeing people as appropriate objects of reactive attitudes, and therefore is situated in the participant view. Strawson thinks that there are both appropriate and inappropriate objects of reactive attitudes (things which are not persons, or persons whose agency is defective are in appropriate), but that with respect to normal adult agents, we can shift between

4 I am helping myself here to the idea that the two viewpoints involve two corresponding types of explanation of action.
5 Strawson 1963:79.

the objective and the participant views (though he thinks we cannot sustain moving away from the participant view). You can, at least temporarily, see someone as a bit of mere nature, and when you do, you won't see them as an appropriate object of resentment.

There are at least three prongs to Strawson's strategy for reconciliation between the optimist and the pessimist. One, he argues that the objective and the participant points of view give fundamentally different, 'profoundly opposed,' kinds of explanations, that both give legitimate explanations, and that they do not rule each other out. Two, he argues that the possibility of the truth of determinism does not threaten the legitimacy of the reactive attitudes. And three, he emphasizes how important and fundamental the reactive attitudes, and the view point from which they make sense, are to us, and how hard it is to imagine giving them up. He appears to take it to follow from this that the framework of reactive attitudes does not require an external justification. On the one hand, this is supposed to satisfy the optimist by accepting the possible truth of determinism, by not seeing this as a threat to moral responsibility, and by not introducing any non-naturalistic, 'panicky' metaphysics. On the other hand, it gives the pessimist back a lot of what she rightly thought was missing from the optimist's position, which sees practices of punishment, moral condemnation and approval as merely ways of regulating behaviour in desirable ways (ways in which people can be "managed or handled or cured or trained"). The optimist and the pessimist both assume that the truth of determinism would mean that we would be forced to view people's actions from the detached, objective view. The optimist thinks that this does not threaten moral responsibility, because we can understand notions of punishment, praise and blame from the objective view, as ways of regulating behaviour. The pessimist rightly (Strawson thinks) sees this as missing out on something important, but we can, in Strawson's view, restore what was missing without denying determinism, or introducing any other metaphysical claim, but rather by grasping the ever present possibility of seeing people from the participant, and not the objective view, and understanding the difference between these ways of seeing people. The optimist is wrong to think that the objective view leaves out nothing important; the pessimist is wrong to think that what it leaves out requires a metaphysical solution. The participant view is fundamental to us, is different from the objective view, and would not be undermined by the truth of determinism.

In arguing for the claim that determinism doesn't rule out 'participant' explanations, Strawson looks at the kinds of considerations that we ordinarily take to mollify resentment: on the one hand, explanations which justify or excuse the action, and on the other, reasons for thinking the agent was acting under abnormal stresses, or is psychologically abnormal, and is therefore not an agent with respect to whom reactive attitudes like resentment are appropriate. The former kinds of considerations remain in the participant point of view, but give reasons for resentment not being warranted by the particular action; the latter shift to a point of view from which resentment does not make sense. Strawson argues that the truth of determinism has nothing to do with either of these types of considerations. In relation to the first group, Strawson argues that determinism is never what we consider when evaluating people's reasons, and what they took themselves to be doing, in order to excuse or justify their actions. He says that no one thinks that the truth of determinism implies, for example, that no one was ignorant of what he was doing, or had good reasons for what he did. In relation to reasons for shifting to the objective view, he argues first that the participant attitude

should give way to the objective attitude in cases of abnormality, but that it cannot be a consequence of determinism that everyone is abnormal. Further, in the case of the abnormal, we adopt the objective view because we see the agent as incapacitated in some way, not because we have been convinced of the truth of determinism. And he argues that the acceptance of determinism couldn't lead to a sustained repudiation of the reactive attitudes in the non-abnormal cases, because our commitment to them is too thoroughgoing and deep-rooted. It is simply not in our nature, he says, to be capable of a sustained objectivity of inter-personal attitude. When we do take up the objective view, it is not because of being convinced of the truth of determinism, and we cannot take seriously the idea of abandoning the participant view because of being convinced of the truth of determinism.

Strawson's argument is subtle, nuanced and hard to pin down. In my view, and as the present volume attests, the text is open to a number of quite different interpretations. As I have already mentioned, there are some similarities between Strawson's strategy and Kant's, in his appeal to two different view points we can take on the world. On the other hand, Strawson's solution is often thought of as a kind of Humean compatibilism.[6] Humean-seeming features of his view include his arguing that determinism doesn't threaten freedom, his giving a central role to emotions in our moral life, and his apparent appeal to descriptive psychological facts—the idea that, because of our natures, we cannot give up the emotional responses which characterize the 'participant' attitude.[7]

Understood in this Humean way, Strawson's argument has seemed to many commentators to be inadequate.[8] The idea that it is not in our nature to give up the participant view does not seem to have the justificatory force we would want it to, to respond to the pessimist. Why couldn't our natures involve responses that are premised on false beliefs, or which see the world incorrectly? Strawson argues that we don't ordinarily consider determinism when evaluating agents' reasons for their actions in order to see whether they could be excused or justified, but the pessimist thinks that our ordinary explanatory practices could be premised on a false assumption. I am going to suggest a strategy for defending Strawson's position which involves appealing to Kant. Strawson would not have thought of what he was doing as Kantian in the way I will suggest, but I will argue that Kant's position is helpful to Strawson, and not foreign to the spirit of Strawson's argument. However, I'm not arguing that the view I'm defending is definitively found in the text—in my view

6 It is not surprising that there are threads of both Hume and Kant in "Freedom and Resentment," since this is also true of the rest of Strawson's work. Strawson develops Kantian anti-sceptical transcendental arguments in *Individuals* and *The Bounds of Sense* but also, in *Scepticism and Naturalism*, argues in more Humean style that certain sceptical worries are idle and do not require response, since they attack beliefs which we cannot give up. While stressing the role of emotions in moral life seems closer to Hume, the emotions Strawson looks at involve the Kantian idea that respect for persons involves seeing them as liable to a minimal demand for reciprocal good will.

7 McKenna and Russell 2008:5.

8 To cite a few examples: Pears argues that Strawson bases his case on the idea that "our whole system of reactive feelings and attitudes is an ineradicable part of our lives", but "We may well wonder how much weight this kind of naturalism can be expected to bear" (Pears 1998:248, 249). Haji argues that "reflection on one's development history can and does affect our affective responses….why can't reflection on the causal effects of determinism on our actions also serve to affect our affective responses" (Haji 2002: 205) Pereboom argues that we can give up reactive attitudes, and can even make imagine human social life without them. (Pereboom 2002, 479) 484, 485. Smilansky (2002) argues that we can't give them up, but that they are premised on an illusion.

Strawson's position in the paper is under-determined by the text, and allows for quite different interpretations.

2
Reactive Attitudes and Forgiveness

Part of the point of Strawson's discussion is to bring out the ways in which certain emotional, attitudinal responses to persons, which he calls reactive attitudes, involve seeing persons as free and responsible for their actions. That these attitudes involve seeing people as responsible is central to Strawson's strategy for dissolving the free will problem. Reactive attitudes are affective ways of seeing another person, in relation to the good will (or otherwise) expressed by them in their actions. Seeing persons as free and responsible is part of the content of these attitudes. When you feel gratitude to someone, you see her as responsible for what she did, and it does not make sense to feel gratitude without this. We can't understand gratitude without some idea of responsibility.[9] I will argue that his analysis is extremely helpful for making sense of forgiveness, which Strawson himself situates in the context of reactive attitudes. The way in which reactive attitudes involve a view of the person, the fact that they are affective, and the fact that they essentially belong to the participant, and not the objective point of view, are all, I argue, important for understanding forgiveness.

I have argued that forgiving involves ceasing to have towards an offender the retributive reactive attitudes which their wrongdoing supports.[10] This might sound like a fancy way of saying that forgiving involves overcoming resentment; I think this is roughly right, but I think what is helpful about Strawson's account is precisely the detail and complexity it adds to how we understand what resentment involves, and therefore what is involved in overcoming it. In my view, what is key to understanding forgiveness is a clear and detailed account of the intentional content of resentment, or, more broadly, of reactive attitudes. This is necessary to give an account of the change in feeling forgiveness involves.

Most philosophical work on the emotions stresses that emotions have intentional content: emotions have objects, and there is a way an emotion presents its object as being. Reactive attitudes are a group of emotional, attitudinal response which have extremely complex intentional content. A large part of what is so useful about Strawson's subtle discussion is how it brings this out. When you resent someone, you see her as having done something you are entitled to expect her not to do. It is already clear that this content is complex. It includes, at least, the idea of responsibility (there must be some sense in which she could have not done the thing that she did), and the idea of a legitimate demand, and therefore has normative content. Strawson does not define reactive attitudes,[11] but his discussion brings out a number of features of their complex content. 1) They are *attitudes*, 2) they are *affective* 3) they are *reactive*, 4)

9 Pereboom (2002) argues that analogues of the reactive attitudes could survive accepting hard incompatibilism, but it is clear that these are analogues, with significantly different content. He says that instead of feeling guilt you could feel sad that you were the agent of wrongdoing, and that the analogue of gratitude would be thankfulness and gladness without praiseworthiness (Pereboom 2002:485)

10 Allais 2008.

11 In replying to Bennett, Strawson says that 'it does not seem to me to matter if a strict definition is not to be had' (Strawson 1980: p. 226).

they are *participant* or non-detached attitudes, and 5) they assume a minimal demand for reciprocal good will from persons. In my view, all these features are necessary to properly understand the intentional content of the retributive emotions that are overcome by forgiving. I comment briefly on each.

1) The significance of the fact that reactive attitudes are *attitudes* is that they have a more complex structure than simpler emotions. Feelings like anger, disgust, delight or joy seem to involve one way of affectively seeing their objects. In contrast, having a contemptuous attitude towards you involves having a disposition to a range of feelings in a range of circumstances: I might right now be enjoying satisfied schadenfreude with respect to one of your setbacks.

2) It is central to Strawson's account that reactive attitudes are affective. My main argument in this section will be that this is crucial to making sense of forgiveness.

3) Not all emotions we are liable to in virtue of our being in relations with other persons are reactive attitudes. Relations with persons can involve emotions which are not responses to the other person's good will, such as frustration with their slowness or delight in their humour. And not all of our affective views of others' wills are *re*active, or responses to what another person has done; trust arguably is often not (though loss of trust may be).

4) Reactive attitudes belong to the participant point of view, and this means that they involve seeing persons from a standpoint of involvement in relationships. This is a complex point. Reactive attitudes are response to something a person has done, more specifically, to the way the other person has expressed their attitude to you. 5) Part of the content this response is seeing the other person as subject to a legitimate demand for a degree of reciprocal good will, as responsible for how they choose to respond to this demand, and as liable to evaluation in the light of how they choose to respond to this demand.[12] The last point is crucial: reactive attitudes see actions as reflecting on the agent.

Darwall explains the content of reactive attitudes in terms of the idea of recognition respect: recognizing that the other person is a person, and recognizing that, as a person, they are subject to a demand for some degree of good will, and have a legitimate claim on our good will.[13] While recognition respect is a requirement of seeing someone as an appropriate object of reactive attitudes, the content of these attitudes must go beyond this, since recognition respect is equally the basis for all reactive attitudes—both gratitude and resentment—and does not distinguish between them. Reactive attitudes do not simply see agents as subject to a legitimate demand, but, further, see them in the light of how they respond to the demand: they involve

12 Something which follows from this, in my view, is that Strawson's account is not best seen as a reductive attempt to characterize responsibility in terms of the having of certain emotional responses, but rather as involving an illuminating circle: on this account, moral responsibility cannot be understood without the reactive attitudes, but equally, the content of reactive attitudes cannot be understood without the idea of moral responsibility. As Snowdon puts it: "Strawson's attitude is that the aim of analysis is to reveal conceptual links and connections, thereby illuminating some features, but that there is no favoured basic level of thought to which the goal is to reduce everything else. One might call that a conception of *relaxed analysis*" (Snowdon, 237).

13 Darwall, "Presidential Address to the Central Division of the American Philosophical Association," 2004.

appraisal respect. Unlike recognition respect, which is unearned, and is simply part of what it is to recognize someone as a person, appraisal respect involves an evaluation of a person's worthiness to be admired or looked down on in some way.[14] Reactive attitudes include appraisal evaluations because they see an action as reflecting on the agent's worthiness in relation to a specific way she has fallen below or risen above legitimate expectations.[15]

There are a number of ways in which, it seems to me, this notion of reactive attitudes is helpful for making sense of forgiveness. In this section, I mention one briefly, then focus in more detail on a second. I discuss a third in the next section. The first problem with which Strawson's notion is helpful is the following worry. The starting point of most accounts is that overcoming anger and resentment is central to forgiveness,[16] but many also think that overcoming resentment is not sufficient for forgiveness, and that not all overcomings of resentment count. I could get rid of anger by, for example, forgetting, or by putting you out of my mind. I may have no anger towards you because I regard you as beneath contempt. If you want me to forgive you, my having no anger towards you in any of these ways would not be what you want. One strategy for dealing with this problem is to add conditions to what counts as forgiveness; I have argued that a number of attempts to add extra conditions don't work.[17] It seems to me that we can avoid the need for extra conditions by having the details of the intentional content of resentment in clearer view, and this is where Strawson's analysis is helpful. The worry about the cases where I get rid of my anger for the sake of my mental hygiene, or forgetting, or putting you out of my mind, is that they don't seem to have the right kind of focus on the wrongdoer; this is not true of a change in reactive attitudes. It is part of the content of reactive attitudes that they involve a way of seeing the other person, so overcoming a negative reactive attitude *is* a change in the way you see the wrongdoer. I can stop feeling angry by simply putting you out of my mind, but I can't change my affective appraisal of you without keeping you in view. If I overcome anger by putting you out of my mind, but still appraise you in the way the wrongdoing supports, I have not forgiven you. If we see forgiveness as overcoming, in specific, the negative reactive attitudes the wrongdoing supports, then we will not need to add further conditions. In addition, this can explain which other feelings can be relevant to forgiveness, such as hurt or disappointment, and will enable us to explain why resentment can fade, but yet you can fail to forgive (you can still have a negative appraisal evaluation on the basis of the action, though you no longer feel it in the specific way that makes it resentment).

The second point with which, in my view, Strawson's account of reactive attitudes is helpful is a feature of forgiveness which has seemed difficult, even paradoxical to a number of philosophers: the problem that the resentment forgiveness overcomes is warranted or justified or appropriate. Why is it a good thing to overcome a way of

14 For example, Michelle Mason argues that shame is a reactive attitude, saying that "To experience shame is to experience oneself as diminished in merited esteem on the ground that one has violated some legitimate ideal of character." (Mason 2010, 417–18).

15 These expectations need not be seen as straightforwardly moral, and may be personal and specific to specific relationships. Feeling hurt by something you did might involve seeing you as failing to consider me in a specific way I expected to you consider me, without seeing this as a failure to meet a general moral demand.

16 Two classic accounts are Jeffrie G. Murphy and Jean Hampton 1988 and Joseph Butler 1913.

17 Allais 2008.

seeing someone which is warranted or justified? How is it even rational? In my view, the fact that reactive attitudes are *affective* attitudes is crucial to making sense of this.

The more attention philosophers have paid to the idea that resentment has intentional content, and that this can be warranted, the more problematic forgiveness has seemed. We start with the idea that wrongdoing warrants resentment: a person has responsibly failed in relation to a legitimate expectation, and negative appraisal evaluations of her are part of seeing this. If the way you see someone is never affected by her actions, you are not really seeing her actions as flowing from her; appraising her in the light of her actions is part of what it is to see them as her actions. We could come to give up resentment by seeing that the act wasn't really her fault: she has an excuse. Or we could come to see that although her action seemed wrong, it was actually justified in the circumstances. Or (in non-serious cases) we could decide to re-evaluate what she's done, and no longer think of it in relation to a demand we hold her to: yes, she was late again, but is being late really so bad? When we excuse, justify or accept what the person has done, we come to a position in which we judge that there is nothing to forgive. If I think that what you did was justified, or that you were not really responsible for it, forgiveness is not at issue: there is nothing to forgive. The possibility of forgiveness comes into play in precisely in relation to unexcused, unjustified, unacceptable wrongdoing, which warrants resentment (or, more generally, negative appraisal evaluations). As Strawson says, "To ask for forgiveness is in part to acknowledge that the attitude displayed in our actions was such as might properly be resented."[18] But if resentment is rationally warranted by what the wrongdoer has done, then it seems that giving it up would involve not seeing the world in the way which the evidence supports. We can put this point in Strawsonian terms. The two kinds of consideration Strawson considers as are relevant to the appropriateness of resentment involve continuing to see someone as an agent, but no longer seeing her as blameworthy or no longer seeing someone as an agent—seeing her from the 'objective' view. The apparent problem with forgiveness is that it involves overcoming resentment without either of these type of considerations coming into play. You remain in the participant point of view, continue to see the wrongdoer as an agent; do not cease to see her as blameworthy, do not cease to see the act as unjustified and unexcused in the way which makes resentment appropriate; yet you overcome resentment.

One strategy philosophers have appealed to to respond to this worry is to see forgiveness as requiring conditions which undermine the warrant for resentment, most centrally, apology and repentance. The idea is that apology and repentance introduce a change in the evidence which warrants the resentment, thereby making it rationally and morally acceptable to give up the resentment.[19] On this kind of conditional account, forgiveness involves giving up resentment which you should not have, because it is no longer appropriate or warranted. Forgiveness becomes a matter of

18 Strawson 1963:76.

19 A very clear example of this strategy is Griswold's demanding conditional account of forgiveness. Griswold says: 'if moderated resentment is still warranted all things considered, then forgiveness is impossible or premature. Forgiveness does not attempt to get rid of warranted resentment. Rather, it follows from the recognition that resentment is no longer warranted. And what would provide that warrant can be nothing other than the right reasons" (Griswold 2007:43). A similar, but less demanding, move is made by Hieronymi, who says: "an articulate account of forgiveness would explain what revision in judgment or change in view would serve to *rationally* undermine justified resentment" (Hieronymi 2001: 535-6).

correctly calibrating your judgements. This, it seems to me, is less than what we want from the notion of forgiveness. Though I cannot argue this here, it seems to me that there are a number of problems with this kind of account. It fails to capture the elective nature of forgiveness;[20] it rules out unconditional forgiveness; it rules out forgiving the dead;[21] it rules out the possibility of foolhardy or mistaken forgiveness (mistaken forgiveness becomes, by definition, not really forgiveness), and it over-intellectualises and over-moralises forgiveness. Rather than argue for this here, I simply want to point out that an alternative strategy emphasizes the differences between feelings and beliefs.

As already noted, most philosophical work on emotions stresses the idea that emotions have intentional content: they have intentional objects, and there is a way an emotion presents its object as being. Some philosophers have thought that feelings are not intentional states, and therefore that the intentionality of emotions must be explained in terms of their having beliefs or desires as components.[22] In my view Strawson's notion of reactive attitudes is best made sense of in terms of an account of emotions which does not assimilate the intentional content of emotions to other intentional states (such as beliefs or desires), but rather allows for the idea that there can be states which are both essentially intentional and essentially feelings. This is not to deny that emotions have relations to straightforwardly cognitive states like beliefs, but simply to deny that we need to reduce them to such states to capture their intentionality. Emotions are not belief states with an add-on of non-intentional feeling; they are feelings which present the world in certain ways. This account of emotions is developed by Peter Goldie and Robert Roberts,[23] amongst others. Other approaches to emotions which link them essentially to intentional content without characterising this content in terms of beliefs are those which understand emotions in terms of evaluative presentations, ways of seeing, concern-based construals, and/or terms of patterns of interpretation and salience.[24] In explaining the idea of essentially intentional feelings, Goldie highlights a number of features of emotions which distinguish them from straightforwardly cognitive states like beliefs: they are distinct in their phenomenology, they are sometimes cognitively impenetrable, they can sometimes be

20 In saying that forgiveness is elective, I do not mean to imply that there cannot be better and worse reasons for forgiving, or to deny that repentance is a standard ground for forgiveness. Rather, the idea is that forgiveness is seldom, if ever, something a wrongdoer is in a position to demand, and that victim's have a good deal of discretion with respect to whether they forgive, such that one can forgive in the absence of repentance, and one can (at least in many cases) refuse to forgive even where there is repentance, without making a moral or rational mistake. I think this is part of our ordinary concept of forgiveness, although I do not argue this here.

21 Of course, whether and in what sense forgiveness is elective, whether unconditional forgiveness is possible, and whether we can make sense of forgiving the dead are contested. But it seems to me that ordinary usage allows these cases, and that an account of forgiveness which can accommodate them is less revisionary than one which can't.

22 This account of emotions influences some readings of Strawson: some commentators seem to think that the intentional content of reactive attitudes must be understood in terms of beliefs they contain. For example, Bennett says that Strawsonian reactive attitudes are not propositional (Bennett 1980: p. 24), Wallace says that Strawson's account does not clearly manage to credit reactive emotions with propositional objects (Wallace 1994: p. 19).

23 Goldie 2000; Roberts 1988.

24 See De Sousa 1979; Goldie 2000; Roberts 1988; Rorty 1980; For example, De Sousa argues that 'emotions can be said to be judgments rather in the way that scientific paradigms might be said to be 'judgements': they are what we see the world 'in terms of.' But they cannot be articulated propositions…[P]aying attention to certain things is a source of reasons, but comes before them' (De Sousa 1979: pp. 138-9).

directly subject to the will (or more subject to the will than belief) and they do not have the same relation to evidence as beliefs have.[25]

Understanding emotions as involving essentially intentional feelings, and stressing the idea that resentment involves emotion, is one important part of what we need to make sense of the idea that the resentment which forgiveness overcomes is warranted. Two ideas that are helpful here are the idea that emotions can be more directly subject to the will than beliefs are (which enables us to make sense of the idea that forgiveness can be a choice), and the idea that they have different relations to the evidence. As Roberts puts it, a rational person has more options with respect to her feelings than with respect to her beliefs.[26] If we understand the intentional content of resentment in terms of warranted beliefs about what an agent has done, or about what her culpable wrongdoing reveals about her, then resentment will be governed by the epistemic norms governing belief formation. This will make it irrational to give up resentment without a change in the evidence which makes it appropriate. In contrast, the norms which govern emotions need not be understood as straightforwardly epistemic.[27] Affective attitudes can be made intelligible, appropriate or fitting by the evidence, without its being the case that they are rationally or epistemically mandated.[28] On this view, the way in which emotions such as resentment are made appropriate by the facts about wrongdoing does not imply that it would be a rational mistake not to resent. Wrongdoing entitles us to resent, but this does not mean that it obliges us to resent.

This provides at least part of what we need to make sense of an account of forgiveness in which it involves giving up resentment which is appropriate, resentment to which we are entitled. On the conditional judgment account, the value of forgiveness lies in the victim sincerely recognizing and giving the offender what is in fact her due. On my account, forgiveness involves giving the wrongdoer *more* than is her due, something to which she is not entitled. Forgiveness involves affectively regarding people as better than their action supports seeing them as being. Rather than a careful weighing of the evidence with a view to working out whether a change in judgment is warranted, it involves a kind of generosity, in which we make a shift in our (affective) view of her character, where this shift in view is neither epistemically mandated nor epistemically forbidden. This is possible because resentment is not an epistemic state

25 Goldie 2000: p. 78. See also Roberts 1988.

26 Roberts 1988:198.

27 The relevant norms include being proportionate, disproportionate, appropriate or intelligible, rather than true. The idea that emotions involve patterns of attention, interpretation and salience is helpful here, because shifts in attention and focus are under-determined by epistemic considerations. Shifts in focus, attention and interpretation are of course constrained by the evidence, and they play a crucial role in the determination of belief, but they are not mandated by the considerations which warrant beliefs. Resentment is not simply a judgment about someone, but a way of focusing on her.

28 Think of optimism and pessimism (I think it is significant that Strawson uses these terms for the opposing positions he discusses). The optimist and the pessimist may have access to the same facts, and neither may be making an epistemic mistake. In a slightly different, but related point, in "Social Morality and Individual Ideal," Strawson discusses the evaluative ideals, visions and pictures that shape our lives, saying that this is "a region in which there are truths which are incompatible with each other. There exist, that is to say, many profound general statements which are capable of capturing the ethical imagination." Strawson thinks that these statements "often take the form of general descriptive statements about man and the world," but "it is wholly futile to think that we could, without destroying their character, systematize these truths into one coherent body of truth….the injunction to see life steadily *and* to see it whole is absurd, for one cannot do both".

which is rationally mandated by the evidence for it, but rather is an affective way of seeing someone, to which the evidence entitles you, but which it does not mandate.

3
Forgiveness and participant attitudes

It seems to me that the discussion so far is not a complete solution to the problem of the rationality of giving up warranted resentment, and that we need to say more to explain what this shift in affectively seeing someone involves—what its content is. In particular, more needs to be said to explain how resentment can have intentional content that correctly sees someone as culpable for wrongdoing and liable to evaluation in the light of it, but yet contains a view of her which is in some sense optional. How can we be *entitled* to see an act as reflecting on the wrongdoer's character unless it *does* reflect on her character? How is it that you see someone, when you do not change your judgment that she has culpably done something she shouldn't have done, yet you somehow disassociate this wrongdoing from her, and no longer let it inform the way you affectively evaluate her. The change in the way you see her does not involve ceasing to judge that she is responsible for the wrong, and it does not involve no longer judging that it was wrong. What does it involve?[29]

Those who think that the idea that emotions are more optional than beliefs is sufficient to explain forgiveness might point out that we can, for example, cognitively recognize relevant sources of danger without necessarily *feeling* fear.[30] It does not follow from the fact that fear is warranted (that there is a genuine source of danger), that there is some moral or rational failing in not having fear, or that not having fear means perceiving the world incorrectly (so long as you have the requisite judgments about sources of fear). Similarly, it might be thought that we can perceive the truths about responsibility without having affective appraisal evaluations of agents. This argument assumes that we can have a judgment containing the relevant cognitive content concerning danger without affect. Even if this is true in the case of fear, it is harder to make sense of in the case of reactive attitudes: it would require having a purely cognitive judgment about the wrongdoer's merited esteem, without having accompanying affect. It is not clear that we can capture such appraisals of agents without affect, and, if we could, continuing to have such a judgment would not seem compatible with having forgiven. Suppose I say to you: 'well, I've stopped feeling angry, but I'll never think of you in the same way again, and I don't trust you.' You will not feel like you have been forgiven. I suggest that the next step in explaining the possibility of forgiveness is to take seriously the idea that reactive attitudes make sense only from the participant point of view: affective appraisals of a person's will are fundamentally different from 'objective' explanations of their actions, and the kind of judgments that are possible without affect are fundamentally different to those which see people as responsible.

---

29 There are a number of answers philosophers have given here, for example, coming to see the person as overall decent. A common problem with these answers, in my view, is that they see the shift forgiveness involves as too general, and insufficiently related to the thing for which you are forgiving someone. One may fail to forgive someone for a specific wrong, while never ceasing to see them as overall decent. With respect to an uncharacteristic wrongdoing, one might continue to judge that the wrongdoer is not generally disposed to act in this way, while still esteeming her a bit differently, and a bit less, in the light of the fact that she was prepared to act in this way on this occasion.

30 Whether this is psychologically realistic is however a serious question. See Damasio 2005.

As I understand Strawson's position, 'objective' explanations involve certain kind of causal explanations—determining causal explanations.[31] When we view people's actions as explainable by determining causes, we are giving 'objective' explanations of their actions. Crucially, these kinds of causes can include psychological states.[32] For a straightforward naturalist and compatibilist, judgments about responsibility simply are judgments about certain kinds of empirical causes—those involving, amongst other things, the agent's beliefs, desires, inclinations, etc. There is therefore no 'profound opposition' between explanations which see the agent as responsible and the 'objective' view. On Strawson's account, however, the content of judgments which see agents as responsible and liable to evaluation in the light of how they have lived up to legitimate demands, are fundamentally different from judgments about empirical causes. Like complex bits of mere nature, we can come to have increasingly detailed information about how a person is likely to act. Just as you could come to think a clock is unreliable in response to its frequently giving you unreliable information about the time; similarly, you can come to have reason to predict that a person is likely to act in particular ways. One could praise or criticise a clock on the grounds of its reliability, but this seems to be fundamentally different from the appraisals that are central to reactive attitudes. Praising a clock or a car for its reliability does not reflect on its *worthiness*, or regard its reliability as an expression of good will.[33] In contrast, as Wallace argues, 'the actions of morally responsible people are thought to reflect specially on them as agents, opening them to a kind of moral appraisal that does more than record a causal connection between them and the consequences of their actions.'[34]

I suggest that Strawson's opposition can be explained in terms of two different ways of thinking about the relation between reasons and actions. On the first account, reasons are understood as sums of beliefs, desires and inclinations that together are causes of action. They are like any other empirical causes in the physical world. On the second account, reasons are not causes. Rather, agents choose to take certain features of the world, which may include their own inclinations, as making certain actions choice-worthy. Agents choose to act for reasons. Citing an agent's reason does not tell you why certain empirical causes, including some that went through her brain, lead to certain outcomes in the world, rather, it makes the action intelligible from her point of view. It picks out what she took to be choice-worthy about the action. These two different ways of understanding the relation between reasons and actions give us a way of understanding the contrast between the explanations of action corresponding to Strawson's 'objective' and 'participant' views. When we see people's actions as caused by psychological states like beliefs and desires, we view them from the objective view. When we see people from the participant view, we understand their reasons as the features of situations they took to be choice worthy, for which they

31 Strawson of course is not arguing that determinism is true, but he is considering whether, if it were true, it would rule out 'participant' explanations of actions.

32 Strawson says that we adopt the objective view to someone when we see them as an object of social policy, "as a subject for what, in a wide range of sense, might be called treatment…to be managed or handled or cured or trained," and he says that psychotherapists see their patients from the objective view.

33 It may be objected that nothing has been said to justify thinking that there is this difference between persons and clocks. Strawson's aim, in my view, is not to provide this kind of justification, but to bring out what is involved in the way we see persons, and how much is at stake in giving this up.

34 Wallace 1994:52. He compares this with praising a painting, where praise and admiration reflect at kind of credit on the artist.

acted, which express their values and attitudes, rather than as the causes of their actions. On the first view, an agent's choosing tells us what her dominant psychological states were: the ones that caused her to act. On the second view, an agent's choosing doesn't tell us something about what caused her to act, it tells us something about what she took to be choice-worthy, or a reason for acting.

In my view, this contrast makes sense of Strawson's claim that 'objective' and 'participant' explanations are profoundly opposed, and that reactive attitudes do not make sense from the objective view. Think about the contrast between praising a reliable clock or car for its good workings, and esteeming a person's good willing. If we see reasons as causes, then what it means to say that an agent's choices reflect her willing is that it tells us about what her dominant beliefs and desires were: the ones which caused her action. This tells us about the causal structure inside her, in a way which is not, in principle, different from the causal structure inside a clock. It fails to capture a difference between the way in which an agent's choosing reflects on her and the way in which a clock's reliability reflects on its inner workings, and therefore fails to capture the way we esteem persons in a special way linked to seeing them as responsible. It is crucial to Strawson's account that there is such a difference. In Strawson's picture, as I understand it, neither recognition respect nor esteem respect are empirical judgments. This is perhaps easier to see with recognition respect: the idea that seeing a person as a person involves seeing her as subject to a legitimate claim for reciprocal good will cannot be cashed out in terms of some empirical property of her that science investigates. Similarly, our evaluations of persons in the light of how they choose to respond to legitimate claims on their good will are not empirical judgements about beliefs, desires and inclinations that caused an action.[35] Participant explanations have an essentially evaluative component: they have built into them the evaluative notion of a legitimate demand, and they involve seeing the agent as liable to a kind of evaluation which is not simply a judgment about what caused her action.

I suggest that this is helpful in making sense of the way in which appraisal evaluations can be more optional than empirical judgments about causes. Suppose that judgments about responsibility were merely judgments about which beliefs and desires caused an action. If this were the case, we would not be able to give an adequate account of the shift that forgiveness involves. It could involve coming to a revised view about an agent's dominant psychological states, either (irrationally) without further evidence, or as a result of her having demonstrated that she has changed. Or it could involve a shift in affect without a corresponding shift in judgment about how to appraise the agent. These both seem to me to be unattractive options. The first makes forgiveness either irrational, or simply a matter of correctly calibrating your judgments about what a person's beliefs, desires and inclinations are like, requiring her to prove that her dominant psychological states are not of the sort which the wrongdoing supports

35 If they were, it would be much harder to make sense of responsibility. As Galen Strawson argues, if we see actions as caused by 'your overall mental makeup'—your desires, beliefs, inclinations, etc, we can't see any one as responsible, since no one is ultimately responsible for their overall mental makeup. He says that "it is your overall mental makeup that leads you to do what you do when you act or deliberate" (G. Strawson 2002, 445) but that "[b]eing the sort of person one is and having the desires and beliefs one has, are ultimately something one cannot control, which cannot be one's fault; it is one's luck." (G. Strawson 2002, 493) This is in keeping with my Strawsonian view, but, on this view, when you see a person as responsible you precisely don't see them as caused to act by their desires, beliefs and inclinations.

thinking they are. This fails to capture the elective nature of forgiveness. The second fails to do justice the intentional content of resentment, as it sees the resentment that forgiveness overcomes as separable from the appraisal evaluations of agents. Both strategies see the appraisals involved in seeing people as responsible in terms of objective explanations. This is precisely what Strawson denies.

An alternative, and in my view more promising approach, is to say that the appraisal evaluations relevant to forgiveness are not judgments about the set of psychological states which caused the action. Rather, they involve seeing someone as something more than, or other than, the totality of her empirical psychological states. When we appraise an agent from the participant point of view, we don't simply see her as a sum of beliefs, desires and inclinations, with respect to which there is a fact of the matter about their dominant causal tendencies, rather, we see her as having an ongoing capacity to choose to respond to value, in a way which is not determined by her beliefs, desires and inclinations. We see an agent's choices not as caused by her inclinations but as expressing her willing or her choosing. This, I will argue, allows for the kind of optionality that is involved in forgiving and enables us to explain how we can rationally appraise the wrongdoer as better than her actions indicate her to be.

Many aspects of this suggestion require explanation: I focus here on two difficulties. The first is whether we can say more to say what responsibility attributions involve, if they do not involve seeing actions as caused by beliefs and desires. The second is to explain why the objective and participant explanations do not exclude each other. I suggest that there is a reading of Kant's account of the free will problem that enables us to develop Strawson's account in a way that begins to answer these questions.

4
The Third Antinomy

The third Antinomy is Kant's attempt to resolve the free will problem in the first *Critique*. Famously, Kant thinks there are equally good arguments for the claim that we have freedom and for the claim that we do not have freedom, and he thinks that considering the arguments on their merits will result in our being in a state of ceaseless vacillation, in which we are simply convinced by which ever side we've been considering more recently (because both sides have convincing arguments).[36] Kant's solution to the problem is, in some respects, unambitious. He thinks that we cannot prove that we have free will (or that we don't have it), and we cannot really understand what it would be to have free will. His less ambitious aim is to show that it is possible that we have it, in particular, that it is not ruled out by what we know about the world in science and metaphysics. On the other hand, there is also a way in which Kant's account seems crazily ambitious. He thinks that something like agent causation is possible, while lots of philosophers think it is incoherent, and he is an incompatibilist, who thinks that determinism would rule out freedom, and who, like hard determinists, thinks that determinism is true, but also thinks he can show that freedom could be true.

Kant's strategy for making this work is to invoke his transcendental idealism: his distinction between things as they are in themselves and things as they appear to us.

36 More precisely (since his arguments for the thesis and antithesis proceed by *reductio*), we will be convinced of the opposing side, because both sides are inadequate and problematic.

Immediately following the publication of the first *Critique*, and for a long time afterwards, Kant's distinction was seen as an extreme metaphysical distinction between non-spatio-temporal, non-sensible objects (noumena) and appearances which exist as ideas in subjects' minds, like Berkelean objects (phenomena).[37] Until relatively recently however, the dominant reading of transcendental idealism rejected this approach, and emphasized instead the idea of two different ways of considering the same objects. So-called deflationary interpretations argue that Kant's distinction is not metaphysical, but rather is an epistemological or methodological distinction between two ways of thinking about things.[38] Many deflationary interpreters deny that Kant is actually committed to there being things in themselves, understood as an aspect of reality of which we cannot have knowledge, and rather think his point is that we cannot avoid the thought of the thing in itself. This two view-points understanding of transcendental idealism is the one most often invoked in discussions of Kant's moral philosophy.

In my view, deflationary interpretations are wrong, because transcendental idealism is a partly metaphysical position, containing a genuine idealism, and a genuine commitment to the existence of an aspect of reality we are unable to cognize. But this need not (and, I think, should not) be understood in terms of the old, extreme metaphysical interpretation. Kant is not a Berkelean idealist or a phenomenalist about appearances, and his commitment to things in themselves is not a commitment to the existence of non-sensible, non-spatio-temporal objects which exist in addition to the spatio-temporal objects of our experience. Kant calls such objects noumena in the positive sense (examples are Leibnizian monads and Cartesian souls), and explicitly denies that his notion of things in themselves should be understood in this way.[39]

The idea that transcendental idealism is partly metaphysical is crucial to how Kant understands freedom in the Third Antinomy (however, I will argue that there is a sense in which his solution is not metaphysical). Kant does not simply talk about different viewpoints we can take on the world; he explains free will, and the conflict between free will and determinism, in terms of the idea of *two different kinds of causality*. He thinks that if the only kind of causality is causality in accordance with laws of nature (his version of determinism), then freedom is not possible. Freedom requires the possibility of a different kind of causality. Causality in accordance with laws of nature involves events which are necessitated to unfold the way they do as a function of previous states of the universe plus the laws. The other kind of causality involves the power to initiate a new causal sequence which is not a necessitated function of previous states of the universe and the laws of nature. Kant thinks that when we see agents as free we see them as having this capacity, and he thinks that transcendental idealism is required to make sense of the possibility of this second kind of causality. Although I think transcendental idealism contains genuine (though non-phenomenalist) idealism, I think the idealism part of it (the mind-dependence of appearances) is not what is crucial to the Third Antinomy. Rather, what is crucial is the idea that there exists an aspect of reality of which we cannot have knowledge: Kant's commitment to things in themselves, and his argument that we cannot have knowledge of their natures, as they are in themselves.

37 This is Strawson's (1966) reading of transcendental idealism.
38 Henry Allison (1983) is a central leader of this shift.
39 I defend this in Allais 2004, 2007, 2010.

Although the third Antinomy is concerned with a metaphysical problem (free will), there is an important sense in which Kant's solution is not straightforwardly metaphysical. His argument in the third Antinomy is part of his attack on transcendent metaphysics: metaphysics which tries to make substantial claims about things which are not possible objects of experience. It is crucial to see that this critique applies just as much to empiricism as it does to rationalism. It is easy to see that the rationalists are doing transcendent metaphysics when they make claims, for example, about monads and Cartesian souls. It is less obvious with the empiricists, because they seem to stick to science, and talk only about the kinds of objects science studies. However, Kant thinks that the empiricists overstep themselves, and pass, without noticing it, into the territory of transcendent metaphysics when they start thinking that science does or could give a complete account of reality. Part of the point of transcendental idealism is to argue that this is not the case: in principle, Kant argues, science cannot be complete, and cannot tell us about the intrinsic nature of reality, the way things are in themselves. Kant thinks that science gives us knowledge only of relational aspects of things (powers) and leaves us ignorant of the intrinsic natures in virtue of which things have the powers they do. One way of understanding Kant's strategy in the third Antinomy, I suggest, is to see him as arguing that it is a mistake to take science as metaphysics: it is in doing this that the empiricist, unwittingly, moves into transcendent metaphysics.

In addition to his argument that we cannot know the intrinsic nature of reality, one of Kant's central concerns in the *Critique* is to establish the limit and status of those metaphysical claims we can know (a metaphysics of experience). Kant thinks that the claim that every thing that happens has a cause which falls under a scientific law (his version of the thesis of determinism) is a synthetic *a priori* metaphysical claim. It is not a truth of logic, and it is not an empirical claim. The status of such claims, and how they could be established, is a problem—the problem to which the whole *Critique* is addressed. Kant thinks that the only way such claims can be established is by being shown to be conditions of the possibility of empirical knowledge.[40] And he thinks he provides a proof that the principle of determinism is such a condition. However, this vindication does not and could not show that the principle applies to, or is true of, everything that exists. Since we can vindicate it only as a condition of the possibility of empirical knowledge, we can know it to be true only of the things of which we can have empirical knowledge. Independently of his view that we cannot know the intrinsic nature of reality, Kant raises an important challenge here. We have no entitlement to assert that what we can know exhausts reality, so we have no basis for thinking that conditions of the possibility of empirical knowledge (such as the principle of determinism) are metaphysical truths about all of mind-independent reality. Of the causes of which we can have empirical knowledge, the causes we can know scientifically, we are entitled to assert that they fall under scientific law. But this gives us no entitlement to assert this is the only kind causality there is, or that everything that happens has a cause that falls under a natural law: it is only of phenomena (things of which we can have empirical knowledge) that we can assert the principle of determinism.[41] This, Kant thinks, is where the space for free will comes in. Since we do not have any justification for asserting the principle of determinism of

40 Kant talks about conditions of the possibility of experience, but by 'experience' he means empirical knowledge. Thus, the determinist principle is not proved with regard to everything of which we could have experience in the sense of consciousness, but something far more specific. 'Phenomena' are the aspect of reality, of which we can have empirical knowledge, which can be known by science.

everything that exists, we are not entitled to say that freedom (the other kind of causality) does not exist. It is possible that there is a second kind of causality.

Kant's solution is unambitious since he does not try to show that we have free will (that there is an alternative kind of causality); he thinks this cannot be shown. From the point of view of what we can know, his argument makes free will a mere epistemic possibility. But this mere possibility needs to be put together with the fact that, Kant thinks, we *do* see ourselves as free and responsible, and this is central and fundamental to the way we think of ourselves. It is, Kant thinks, at the basis of morality and responsibility. He thinks that in recognizing moral reasons—seeing that there are things we ought to do—and in seeing people as responsible for how they respond to moral oughts, we see ourselves as having a capacity to initiate actions that are not a function of previous states of the universe. Since, Kant thinks, this central way we see ourselves is not ruled out by what we know about the world in science and metaphysics, we are entitled to continue with it. Rather than trying to prove that we are free and responsible (which, Kant thinks, cannot be done), we start with the fact that we do see ourselves as free and responsible. Metaphysics and science, Kant thinks, seem to threaten this way we see ourselves; his aim is simply to ward off this threat. He argues that we mistakenly think that science rules out the possibility of freedom when we mistakenly take science for metaphysics.

It is important that Kant thinks that the only kind of proper causal explanations we can give are empirical causal explanations, and he thinks that when we explain an agent's action by appealing to her reasons we are not giving an empirical, causal explanation. Empirical causal explanations tell us why something had to follow, given previous conditions and laws; we have nothing comparable to this with respect to explanations of actions. Giving the agent's reasons tells us what she took as choice-worthy, but it does not tell us that, given the reasons she had, she had to act as she did, and it does not tell us why she chose as she did. She took this as a reason for action; she could have chosen not to act for this reason, and there is no further explanation of why she acted as she did than that this is what she chose. This means that there is a sense in which agent's choices cannot be explained. However, Kant thinks although we don't have proper causal explanations here, we still have the idea of a kind of causality in which an agent has a capacity to choose for reasons, and to initiate a new causal chain which is not a function of previous states of the universe.

Although Kant thinks that we cannot understand or explain the kind of causality involved in acts of free choosing, we can say a bit about it. First, while we cannot give further causal explanations of the sort which tell us why something had to happen, we have an idea of causation in the following sense: we see the agent as the initiator of

41 There are of course many complications as to how to understand this solution, bound up with the controversial question of how to understand Kant's transcendental idealism. It is unclear how we should understand the idea that determinism is true of phenomena, but freedom is possible in things in themselves, and a common response is to say that it is questionable why we should care about such freedom, if it makes no difference to the empirical world. I cannot resolve this complicated question here; I simply comment that, in my view, it is crucial to see that phenomena, on Kant's account, are not ontologically self-sufficient and complete. This means that phenomena are not a complete, independent, determined aspect of reality. Rather, what is true of phenomena, including what laws there are, is partly a function of the way things are in themselves, so what is true of things in themselves does make a difference to the empirial world.

her action.[42] Second, we can characterize the way we see agents when we attribute responsibility to them negatively, by saying that we do not seeing the action as a determined result of the agent's psychological states. Third, we can say something positive to explain the relevant capacity to choose (the capacity for free action), though what we can say is not metaphysical but moral: it involves the idea of having the capacity to recognize higher-order rational constraints on thinking about reasons for action.[43]

6
Strawson and Kant

Strawson does not see his strategy as Kantian, but Strawson, in *The Bounds of Sense*, read transcendental idealism in terms of the extreme metaphysical interpretation. Once we set this aside, there are some similarities in their strategies. A central similarity is that neither of them argues for the overall defensibility of responsibility attributions: they both take as fundamental, as a starting point, the idea that we do see people as responsible. Both think that we cannot make sense of responsibility within the kind of explanations offered by science (the objective view, or deterministic causality). Both think that we mistakenly think the truth of determinism would threaten our responsibility attributions, both aim to ward off the threat, and both aim to do this without proving any contrary metaphysical thesis. Both take determinism and scientific explanation seriously, yet think there is a question about the status of the principle of determinism.[44]

Like Kant, Strawson thinks that seeing ourselves and each other as responsible and free is a central and fundamental part of our understanding of ourselves, and he emphasizes how hard it is to think of giving it up. On the Humean reading of Strawson, this is an appeal to descriptive psychology: an insistence that it is simply not in our nature to cease to have reactive attitudes. As noted at the beginning, this does not seem like much of an argument against the pessimist, who thinks that our natures could simply involve false views of the world. In contrast, my view is that Strawson's point is not simply an appeal to descriptive psychology, and it is not meant to be an argument for the claim that our responsibility attributions are warranted or

---

42 Humean compatibilists also have an account of agents as initiators of their actions: agents initiate their actions when their actions are caused by their psychological states in the right way. Kant's account of initiation involves the very strong idea of starting a new causal chain that is not a function of previous states of the universe.

43 In particular, the central rational constraint on thinking about reasons for acting is given by the recognition of other persons as having the capacity to act for reasons and to recognize higher-order rational constraints on doing so—the categorical imperative. Of course, compatibilists can, and do, also appeal to a role for higher order beliefs. However, these will play a different role if they are understood as causes. In Kant's account, having the capacity to initiate action requires the capacity to recognise normative reasons for action, and, the other way round, having the capacity to recognise normative reasons requires having a particularly causal capacity. He thinks that there is an essential relation between the idea of causality and the idea of law; empirical causality takes place in accordance with laws of nature, the second kind of causality involves the capacity to recognise the moral law.

44 Kant thinks that as a synthetic *a priori* presupposition of science, rather than something that could be empirically proved, we have no justification for asserting it as anything other than a condition of the possibility of empirical knowledge. Strawson says that he is closest to those philosophers who do not understand what the determinist hypothesis is, that determinism is "at present no more than a formal conjecture. We point inarticulately at the total set of antecedent conditions of an action and we guess that if they recurred in every detail, the action would be repeated" (Response to Pears 1998, 253).

couldn't mis-represent the world, but rather an attempt to show how important they are to us, and to bring out what they involve. At least as important as his claim that we are unlikely to be able to give these attitudes up is his emphasis on how "how much we actually mind, how much it matters to us whether the actions of other people—and particularly *some* other people—reflect attitudes towards us of goodwill, affection or esteem on the one hand, or contempt, indifference or malevolence on the other."[45] These appeals, it seems to me, are not meant to be a proof that our responsibility attributions couldn't be wrong; rather, they are meant to persuade you to think yourself into the 'participant' viewpoint, and to see how central it is to the way you see the world.[46] When Strawson insists that our commitment to ordinary inter-personal attitudes is part of the framework of human life, and not something which can come up for review, the idea is not that it is because we naturally have certain attitudes that questions of the justification of these attitudes cannot be raised, but rather that the ways of seeing people these attitudes depend on is as well established as anything else we believe, and neither needs nor could get justification from science. Strawson insists that responsibility attributions are part of 'the facts as we know them.' If we start with a Humean reading, and assume that the facts as we know them are the scientific or empirical facts, then we will read Strawson as saying that responsibility attributions are part of empirical facts, and that this is how they are compatible with determinism. This, it seems to me, cannot be right, since he stresses that we cannot make sense of responsibility attributions from the point of view of objective explanation. Rather, I think his point is that the facts science picks out are not the only facts: like Kant, he does not take science for metaphysics.

The biggest difference between Kant and Strawson is that Kant thinks the idea of freedom, and attributions of responsibility, involve the thought of a second kind of causality.[47] Kant thinks that responsibility attributions involve seeing agents as initiating their actions, and that this means seeing the action as not being a determined function of previous states of the universe plus the laws of nature. Strawson, I think, would class this aspect of Kant's account together with the panicky metaphysics of libertarianism. But I think that Kant's account is less metaphysical than it appears, and that the idea of a different kind of causality is helpful to Strawson's case. Because Kant appeals to a metaphysical theory (transcendental idealism), thinks that freedom involves a different kind of causality, and seems to think that it involves causation by noumenal selves, it is easy to think that his account of freedom is metaphysical. But this is not the only way of understanding it. The Dialectic is Kant's *critique* of transcendent metaphysics, and he argues that freedom is one of the transcendent metaphysical notions of which we cannot have knowledge. He appeals to the possibility of a different kind of causality but he does not give any account of how this works, and he thinks that attempting to give such an account would be a mistake.

---

45 Strawson 1963:76. Similarly, he says "I want to insist on the very great importance that we attach to the attitudes and intentions towards us of other human beings" (Strawson 1963:75).

46 This is not a merely pragmatic argument. "Strawson maintains that the unavoidability of the system makes the question of its justifiability an issue of no practical importance" (Pears 1998, 250)

47 There are of course differences between Kant and Strawson with respect to the way they understand the status of determinism. Kant's views about determinism are both stronger and weaker than Strawson's. We could put Kant's position crudely by saying he thinks determinism both is and isn't true: true of the 'phenomenal' world,' but not of the 'noumenal world', whereas Strawson opens his paper saying that he identifies most with those philosophers who say they do not know what the thesis of determinism is, doesn't argue that determinism is true, and clearly thinks that it is a mistake to think we are in a position to assert its truth.

This means, I think, that Kant's account both is and is not an agent causation account. It's not an agent causation account to the extent that such an account usually attempts to explain an alternative kind of causation, and how it relates to the empirical causal chain. Kant thinks this cannot be done. We want to know how the agent's initiating her action relates to the empirical causal chain (Kant thinks that we cannot avoid trying to answer unanswerable metaphysical questions), but this question cannot be answered. When we try to answer it, we get tangled up. When we try to explain agent causation, the temptation is to think of it as involving causes which are somehow both part of the empirical causal sequence and not part of it.[48] We are looking for something that will explain an agent's actions, sort of in the way that empirical causal explanation does (will show us why the event happened), but also sort of not (the explanation of why it happened won't show that it *had* to happen). Kant thinks that empirical causal explanation is the only kind of proper, knowledge-involving explanation we have of events; there are no further causal explanations accessible to us. However, he thinks it would be a mistake (though a natural one) to assume on the basis of this the transcendent metaphysical claim that this is the only kind of causation there is. This empiricist assumption makes the other kind of causality seem even more mysterious, as it leads to such thoughts as that an action is either determined or is random,[49] either caused by my empirical psychological states and situation or by nothing. Kant thinks that a live alternative is that actions are caused by agents' acts of choosing, and that although we cannot give explanations of this that satisfy us (the only causal explanations we can give involve scientific causation), the idea that agents do initiate actions is an ever present fundamental part of the way we see the world. He thinks that if everything that happened were fully explained by scientific law, there would be no room for freedom; but the idea that everything that happens is fully explained by scientific law is not something science could establish, and not something we can establish as a metaphysical principle (because the only legitimate metaphysical principles are established as conditions of the possibility of empirical knowledge).

Both Kant and Strawson think that scientific explanations which invoke determining causes are not the only legitimate kind of explanation, and that they do not rule out responsibility attributions. I think Kant has more to say in defence of this claim, and that what he has to say is helpful to Strawson's position. On Kant's view, claims like the causal closure of the physical, or the assertion that everything that happens has a cause that falls under a scientific law, are, transcendent metaphysical claims, which cannot be known to be true. Further, Kant's idea that responsibility attributions involve the idea of an alternative kind of causalityseems to me to be a helpful way of

48 Robert Kane explains "Since agents had to be able to act or act otherwise, given exactly the same prior psychological and physical history…some "extra (or special) factors" had to be introduced to explain how and why agents acted as they did. These extra or special factors postulated by libertarians have been various. They have postulated noumenal selves (Kant) or immaterial egos (Cartesian dualists) or "transempirical power centres" that intervene in the brain (Nobel physiologist Sir John Eccles)" (Kane 2002, 415). In my view, Kant thinks that we cannot give an explanation of any extra factor, and we cannot explain why agents choose as they do; however, he thinks that we have the thought of agents being able to initiate actions which are not a determined function of previous states of the universe, and that this thought is part of our view of agents as responsible for their actions.

49 See Van Inwagen 2002:168. Thinking that our actions are either empirically caused (explainable by causes which fall under scientific laws) or random, or caused by nothing, clearly does not get more responsibility or freedom. However, we are not forced to think that this exhaust the options, because neither science nor metaphysics can establish that causation according to scientific law is the only kind of causation there is.

making sense of the content of the responsibility attributions involved in participant explanations of actions. The idea is that these explanations involve seeing agents as initiating their actions, that this is different from seeing their actions as caused by their empirical psychological states, and that we have no basis for excluding the possibility of this kind of causality. If there were not a kind of causality involved, participant explanations would not give us a way of seeing the agent as responsible for her choice; if it were not a different kind of causality, they would not give us fundamentally different explanations from objective explanations.

Strawson's position is often seen as naturalist, and transcendental idealism is not a position that most people think of as naturalist. However, there is a clear sense in which it is. Kant rejects super-natural or non-natural explanations, and argues that the only legitimate theoretical knowledge we have is either empirical knowledge, or knowledge of the conditions of the possibility of empirical knowledge. Naturalism is associated with empiricism, and Kant certainly gives a central place to empirical science. However, he thinks that there is a way in which empiricism can be come non-naturalist: when it claims that science has the capacity to explain everything, it makes a transcendent metaphysical claim. Kant's naturalism involves limiting empiricism.

7
Freedom and Forgiveness

Our problem with forgiveness was making sense of the content of the appraisal evaluations reactive attitudes involve in a way which allows us to see them as warranted yet in some way optional. I suggest that Strawson's idea that the participant explanations are fundamentally different from objective explanations is helpful for understanding this, when we understand it in the way sketched above. When we look at an agent's actions from the participant point of view, we see them as reflecting on her in a specific way: reflecting on her willing or her choosing. We do not simply see her as an empirical character consisting of a sum of beliefs, desires and inclinations which will cause her to act in certain ways. If we did, there would be a fact of the matter about the state of her character, and the only optionality with respect to our judgments of her would be a function of epistemic uncertainty.[50] This, it seems to me, does not leave room for forgiveness. We do not just want to say that forgiveness is possible because although the evidence is that I am entitled to resent you, I could be wrong, we want to make sense of giving up resentment to which you are entitled. The person who is sincerely asking for forgiveness is not saying: you could be wrong about being entitled to resent me; she is saying, I know you are entitled to resent me, but please forgive me. According to the epistemic humility view, there is a correct view of the agent, a fact of the matter about how her character really is, but because we don't know it for sure, we have some flexibility in how we choose to see her. The strategy sees our view of the agent as remaining within the space of objective judgments about action and or character, and this seems to me to be a mistake. I think it misses something crucial about reactive attitudes.

50 This seems to be roughly what Allais (2008) appeals to; it does not seem to me satisfactory.

Consider a parallel case of trust. Here I draw on the accounts of Jones[51] and Baier,[52] who see trust as an affective attitude. On this view, although you should not trust in ways which are not sensitive to the evidence about the way the world is, trust is not simply a belief or cognitive judgment about risk.[53] If 'objective' explanations gave an exhaustive account of reality, there would simply be a fact of the matter about how reliable someone's beliefs, desires and inclinations would cause her to be, and trust would be a calculated risk or gamble taken under conditions of epistemic uncertainty. The view sketched above enables us to give an alternative account. From the participant point of view, we see agents as having an ongoing capacity to choose well, and when we trust them, we have an optimistic attitude towards their goodwill towards us.[54]

From the participant view, we see the agent as something more than, or other than, her empirical psychological states. She has a certain causal capacity: an ongoing capacity to choose to respond to value. I think that this view enables us to make better sense of appraisal evaluations, and the difference between them and the way we see objects like cars and clocks. An agent's choices reflect on her, but not because they tell us what her internal workings/causes are; they reflect her willing or her choosing. On my account, forgiving someone involves seeing them as better than their wrongdoing indicates them to be. This is a possibility, because she can be better than her actions supports seeing her as being. The point is not that forgiving involves merely seeing her as *capable* of acting differently: this is involved in recognition respect, and is the basis for all the reactive attitudes, so cannot distinguish between their content. Forgiveness goes beyond merely seeing someone as having the capacity to act better than she did; it involves seeing her as being better.[55] Forgiving does not involve a new judgment about the totality of an agent's beliefs and desires, but an optimistic evaluation of her willing, going forward. It involves choosing to evaluate her as someone who will respond better, and this is not irrational because she can be this.

Strawson says that if we sufficiently, radically, modify the view of the optimist, his view is the right one. What the optimist has right, he says, is that "(1) the facts as we know them do not show determinism to be false; (2) the facts as we know them supply and adequate basis for the concepts and practices which the pessimist feels to be imperiled by the possibility of determinism's truth."[56] What he says is wrong about the optimist's position is that it gives "an inadequate account of the facts as we know them, and of how they constitute an adequate basis for the problematic concepts and practices."[57] What is inadequate in the optimist's position is that he thinks the facts science picks out exhaust the facts as we know them, and he tries to find a basis for our practices of moral praise and blame within the context of the kinds of causal facts

51 Jones 1996.

52 Baier 1997. See also Becker 1996.

53 As Becker argues, "either I can compute the risk that what you say will be incorrect or I can't. If I can, then what more do I need…Nor is it clear that credulity would be a useful thing in cases in which I cannot compute the risk" (Becker 1996:47).

54 Baier 1997:271. "To trust is neither quite to believe something about the trusted nor necessarily to feel any emotion towards them—but to have a belief-informed and action influencing attitude."

55 If it were an empirical judgment about how her internal states are likely to make her act this would be either unrealistic and irrational (where there is no change in the evidence), or epistemically obligatory (where there is a change in the evidence).

56 Strawson 1963: 73.

57 Strawson 1963: 73.

science picks out. Strawson says that "in philosophy, though it also is a theoretical study, we are to take account of the facts in *all* their bearings; we are not to suppose that we are required, or permitted, as philosophers, to regard ourselves, as human beings, as detached from the attitudes which, as scientists, we study with detachment"[58]

One way of interpreting Strawson's solution is as a Humean kind of compatibilism. Within this account we could say either that participant explanations are part of the empirical explanations, and this is why they are compatible with determinism, or that participant explanations involve no kind of causality at all; they are incommensurable with causal explanations. This position seems to me to be unsatisfactory. The idea that attributions of responsibility are a part of our nature that we cannot give up and *therefore* don't need justification seems a weak and question-begging response to the pessimist. The position is hard to reconcile with Strawson's claim that he is not sure that he knows what the thesis of determinism is. It doesn't make any sense of the specific content Strawson attributes to reactive attitudes, and the fact that they are profoundly different from objective explanations. It does not explain the crucial feature of reactive attitudes: that they involve seeing agents as responsible for their actions in a way which is fundamentally different to seeing bits of mere nature as having determining causes.

I have suggested an alternative to Humean compatibilism, which is a Kantian compatibilist interpretation of Strawson. On this view, the central claims Strawson makes are: 1) Empirical causal explanations involving determining causes are the only causal explanations we can give. 2) Attributions of responsibility are fundamental to the way we see ourselves and to our lives. 3) Seeing agent's actions as responsible and free is fundamentally different from seeing them as caused by psychological states. This position seems to me to be strengthened by the addition of three points from Kant. 1) His argument that we have no basis for asserting that determinism is true of everything that exists (or that causality according to natural law is the only kind of causality); 2) the thought that attributions of responsibility involve seeing agent's as initiators of their actions in a way which is different to seeing actions as having determining causes, and 3) the claim that this is not ruled out by the way we think about the world in science.

In support of my Kantian Strawson, I close with something he says of himself: "I have no religious beliefs. When asked whether I believe in God, I am obliged to answer 'No'; I have difficulty with the concept. But I am sometimes tempted to add that I believe in grace—a quality which eludes precise description, but is sometimes manifested in the words and actions of human beings."[59]

References:

Allison, H., *Kant's Transcendental Idealism*, New Haven and London: Yale University Press, 1983
Allais, L., 'Wiping the slate clean: The Heart of Forgiveness,' *Philosophy and Public Affairs*, 2008, 36(1) pp 33–68.
Allais, L., 'Dissolving Reactive Attitudes: Forgiving and Understanding,' *The South*

58 Strawson 1963: 93.

59 Strawson 1998.

African Journal of Philosophy, 2008, 27, pp 1–23.
Allais, L., 'Kant's One World,' *The British Journal for the History of Philosophy*, Vol. 12, No. 4, 2004.
Allais, L., 'Kant's Idealism and the Secondary Quality Analogy,' *Journal of the History of Philosophy*, vol. 45, no. 3, 2007,
Allais, L., Transcendental Idealism and Metaphysics,' *Kantian Yearbook*, 2, 2010, pp 1–31. pp 459-84.
Bennett, J.. 'Accountability,' in *Philosophical Subjects*, Z. Van Straaten (ed.), Oxford: Clarendon Press,1980.Brown, Clifford, *Peter Strawson*, McGill: Queens University Press, 2006
Butler, Joseph, "Upon Resentment," in *Fifteen Sermons Preached at the Rolls Cathedral*, London: Macmillan and Co., 1913.
Damasio, Antonio, *Descartes' Error: Emotion, Reason and the Human Brain*, Penguin, 2005
S. Darwall, "Presidential Address to the Central Division of the American Philosophical Association," 2004, hhtp://www-personal.umich.edu/ sdarwall/.
de Sousa, R. 1980. 'The Rationality of Emotions' in Rorty, A.O. (ed.), *Explaining Emotions*. Berkeley: University of California Press.
Goldie, P. 2000. *The Emotions*, Oxford: Clarendon Press.
Griswold, C.L. 2007. *Forgiveness: A Philosophical Exploration*, New York: Cambridge University Press.
Haji, Ishtiyaque. 2002. "Compatibilist Views of Freedom and Responsibility." In *The Oxford Handbook of Free Will*, 202–228. Oxford: Oxford University Press.
Hieronymi, P. 2002. 'Articulating an Uncompromising Forgiveness,' *Philosophy and Phenomenological Research*, 62(3), 529–555.
Karen Jones, 'Trust as an Affective Attitude,' *Ethics*, 107(1) (1996), pp 4–25.
Kane, Robert. 2002. "Some Neglected Pathways in the Free Will Labyrinth." In *The Oxford Handbook of Free Will*, 406–437. Oxford: Oxford University Press.
Mason M (2010) On shamelessness. Phil Papers 39(3):401–425.
McKenna, Michael, and Paul Russell. 2008. "Perspectives on P. F. Strawson's 'Freedom and Resentment'." In *Free Will and Reactive Attitudes: Perspectives on P. F. Strawson's "Freedom and Resentment"*, 1–17. Ashgate.
Murphy, Jeffrie G. and Hampton, Jean, *Forgiveness and Mercy*, Cambridge: Cambridge University Press, 1988.
Pears, David, "Strawson on Freedom and Resentment," in *The Philosophy of P.F. Strawson*, ed. Lewis Edwin Hahn, 244–258. The Library of Living Philosophers. Chicago and Lasalle, Illinois: Open Court, 1998.
Pereboom, Derek. 2002. "Living Without Free Will: The Case for Hard Incompatiblism." In *The Oxford Handbook of Free Will*, 475–488. Oxford: Oxford University Press.
R. C. Roberts, "What and Emotions is: A Sketch," *The Philosophical Review*, 97(2) (1988), pp183–209
Rorty, A.O. 1980. 'Explaining Emotions' in Rorty, A.O. (ed.) *Explaining Emotions*, Berkeley: University of California Press.
Russell, Paul. "Strawson's Way of Naturalizing Responsibility." *Ethics* 102 (2): 287–302.
Smilansky, Saul. 2002. "Free Will, Fundamental Dualism, and the Centrality of Illusion." In *The Oxford Handbook of Free Will*, 487–505. Oxford: Oxford University Press.

Snowdon, Paul. “Biographical Memoirs of Fellows: P. F. Strawson.” *Proceedings of the British Academy* 150: 221–244.
Strawson, Galen. 2002. “The Bounds of Freedom.” In *The Oxford Handbook of Free Will*, 441–460. Oxford: Oxford University Press.
Strawson, P.F. 1998. “Intellectual Autobiography.” In *The Philosophy of P. F. Strawson*, ed. Lewis Edwin Hahn, 3–21. The Library of Living Philosophers. Chicago and Lasalle, Illinois: Open Court.
Strawson, P.F.. ‘Freedom and Resentment’, in G. Watson (ed.), *Free Will*, Ox- ford: Oxford University Press, 2004.
Strawson, P. F., *The Bounds of Sense*, London: Methuen & Co. Ltd., 1966.
Strawson, P.F. 1980. ‘Reply to Ayer and Bennett’, in *Philosophical Subjects*, Van Straaten, Z. (ed.), Oxford: Clarendon Press.
Strawson, P. F., "Reply to Pears, in *The Philosophy of P.F. Strawson*, ed. Lewis Edwin Hahn, 244–258. The Library of Living Philosophers. Chicago and Lasalle, Illinois: Open Court.
Van Inwagen, Peter. 2002. “Free Will Remains a Mystery.” In *The Oxford Handbook of Free Will*, ed. Robert Kane, 158–177. Oxford: Oxford University Press.
Wallace, R.J. 1994. *Responsibility and the Moral Sentiments*, Cambridge, Massachusetts: Harvard University Press.