Co-reactive attitudes and the making of moral community

Final MS, forthcoming in In *Emotions, Imagination and Moral Reasoning*, eds., C. MacKenzie & R. Langdon. Macquarie monographs in Cognitive Science. Psychology Press, 2010.

Victoria McGeer
vmcgeer@princeton.edu

Abstract

According to P. F. Strawson, the concepts and practices of "holding responsible", as animated by reactive attitudes and emotions, do not presuppose libertarian free will but what I call co-reactivity: a sensitivity to the scaffolding structure of reactive emotions that is displayed by most human beings most of the time. Many contemporary cognitive theorists, while paying deference to Strawson, have reverted to the idea that a presumption of libertarian free will is essential to reactive practice. Some treat this presumption as a hopeless error, others as a necessary illusion. This divide between Strawsonians and non-Strawsonians has important research implications for cognitive psychology; but more important still, it has great significance for the theory and practice of corrective justice. The hopeless-error theorists will be drawn to a crude consequentialist view of punishment purged of individual blame, the necessary-illusion theorists to an equally crude retributivist view. By contrast, those of a Strawsonian bent should find themselves drawn to a novel restorative vision which pays due deference to the natural kinematics of reactive emotions.

Introduction

The term "reactive attitudes" was introduced by P.F. Strawson (1974) in his paper "Freedom and Resentment", rightly viewed as one of the most important and revolutionary contributions to the free will debate in contemporary philosophical discourse. Reactive attitudes, in Strawson's terminology, are special emotion-laden responses to which human beings are naturally prone in their interactions with one another. They encompass emotions such as (this is Strawson's list): gratitude and resentment; hurt feelings; indignation and approbation; shame and guilt; remorse and

forgiveness; certain kinds of pride; and, certain kinds of love. What makes reactive attitudes special is that they express both a sensitivity to how people are regarded and treated by one another in the context of their interactions, and a normative demand that such treatment and regard reflect a basic stance of good will, modulated to suit the kinds of interactions in question (e.g., as between family members, friends, relative strangers, and so on). As Strawson says, we care enormously whether people manifest good will, affection or esteem in their interactions with others, or if they express contempt, indifference or malevolence. And we care whether we ourselves are the recipients of such treatment or whether others are (e.g. we might feel indignation when someone else is treated badly). We even care when others are the recipients and we are the perpetrators, actual or prospective (i.e., we are prone to feel shame and guilt).

Reactive attitudes form an important subset of our moral emotions; but even given their rich variety, they do not encompass the entire range of our moral sentiments. For instance, we might feel compassion or pity for those we think to be appropriate targets of moral regard. But the reactive emotions – Strawson mentions gratitude and resentment as exemplary instances – are felt only towards those who we think meet more stringent conditions. Such individuals must be appropriate targets of moral regard to be sure; but they must also be capable of showing moral regard in return. It is this capacity of showing moral regard that we take to be critical for responsible agency. So reactive emotions are those emotions we think it appropriate to feel only towards "responsible agents": agents that we consequently hold responsible for their actions and attitudes; not just by responding to them with reactive emotions, but also by engaging in activities that we take to be fitting in light of our reactive emotions, activities such as praising and blaming, punishing and rewarding.

Strawson's paper does a masterful job of laying bare the shape and structure of our reactive attitudes: reminding us of their variety, of the significance they have for us, of the sorts of conditions that lead us to modify or suspend such attitudes, and of their critical role in making and sustaining moral community. But Strawson's purpose in doing this rich forensic work is ostensively a rather arcane philosophical one: it is to

show that metaphysical discussions about whether or not human beings possess a contra-causal or "libertarian" free will are simply irrelevant to the justification of our concepts and practices of "holding responsible"; concepts and practices that are embodied in the complex web of our reactive exchanges.

In light of this, it's worth asking whether Strawson is rightly viewed as a staunch "compatibilist": is he committed to the view that our concepts and practices of holding responsible are compatible with the metaphysical thesis of determinism, according to which every event, including every human actions, is entirely determined by the prior physical state of the universe in accordance with natural law? Characterizing Strawson's position as compatibilist, though not incorrect, might well be misleading. Certainly Strawson rejects "incompatibilism" – the view that our concepts and practices of holding responsible cannot be justified if determinism is true. In other words, he rejects the thesis that maintaining such concepts and practices requires a belief in libertarian (contra-causal) free will. So if the rejection of incompatibilism entails the acceptance of compatibilism, Strawson is a compatibilist. But his is not a justificatory compatibilism. That is to say, he does not regard the concepts and practices of holding responsible as being justified only if they can be squared with the metaphysical thesis of determinism. Strawson's view, once again, is that any such metaphysical justificatory project (whether incompatibilist or compatibilist) is deeply misguided. As he says:

> Inside the general structure or web of [reactive] attitudes and feelings … there is endless room for modification, redirection, criticism and justification. But questions of justification are internal to the structure or relate to modifications internal to it. The existence of the general framework of attitudes itself is something we are given with the fact of human society. As a whole, it neither calls for, nor *permits*, an external 'rational' justification [i.e. in terms of either a libertarian or a deterministic metaphysics]
>
> (Strawson 1974, p. 23; my emphasis).

Why is this issue of justification so important? One answer – perhaps Strawson's own – flows from a disinterested concern with the philosophical enterprise: if philosophers could only get this right, they could liberate themselves from the deep and tired grooves of a pointless metaphysical debate, with compatibilists on the one side urging that causal determinism is no threat to our ordinary practices of holding responsible, and incompatibilists, on the other side, arguing that it is. But, to my mind, there is a far more significant worry lurking in the shadows of various remarks that Strawson makes – one that he never explicitly develops, but which ought to command more attention. It is that certain practical dangers are likely to flow from falling prey to a mistaken belief that metaphysical theses are relevant to the justification of our practices of holding responsible – dangers that threaten the very practices themselves. Of course, there's a bit of irony here. The free will debate is generally motivated by the thought that we cannot hold on to our ordinary concepts and practices of holding responsible without making them metaphysically acceptable: without squaring them with one or another metaphysical picture of human choice. But, in my view, the deeper significance of Strawson's work is to show that these concepts and practices may be vulnerable to damage and distortion precisely as a consequence of seeking to square them with a metaphysical picture of human choice, hence, such metaphysical inclinations are important to nip in the bud before they can flower into noxious, though possibly seductive, practical recommendations.

In this paper, I aim to build on Strawson's insights in a novel context and to novel purpose. In the first philosophical section, I lay out two distinctive theses I derive from Strawson's work indicating why I think they are attractive. In the second more interdisciplinary section, I show that both of these theses have been rejected by a new wave of cognitive research that otherwise takes its cue from Strawson in focussing on reactive attitudes as the key to understanding our folk concepts and practices of holding responsible. And then in the third (brief and schematic) criminal justice section, I explore the practical implications of these rival approaches (Strawsonian and non-Strawsonian) for thinking about our institutions of corrective justice. My aim here is to show that, while the new wave of cognitive research leaves us with the old dichotomy under which

"just punishment" is cast either as strategic conditioning or rigorist retribution, Strawson's approach points us in a quite novel direction, one that I associate with the relatively new movement of restorative justice.

Two Strawsonian theses

The thesis that has attracted most philosophical attention in Strawson's work is one he explicitly defends. I will call it the *metaphysical non-commitment thesis*. According to this thesis, the concepts and practices of responsibility, as embodied in our reactive exchanges, do not presuppose anything so metaphysically demanding as libertarian (or contra-causal) free will. Hence, the truth or falsity of determinism is simply irrelevant to the coherence of these concepts and practices. The metaphysical non-commitment thesis is thus a thesis about what sort of property we must be tracking in one another in order for our reactive attitudes to be properly targeted – i.e., targeted on agents who are not justly exempted (as Strawson says) from our practices of holding responsible.[1] And Strawson's claim is that this property has nothing whatsoever to do with the metaphysical underpinnings of human choice. So what then is this property? Strawson's discussion suggests both a negative and a positive point.

The negative point is that this property cannot consist in having and/or exercising a libertarian "free will". Libertarians defend the view that people are only appropriately held responsible for what they do, if they could have done otherwise right up to the moment they acted. But the only way that they could have done otherwise in this extreme sense is if their choice was not determined by any prior events, but proceeded instead from the free exercise of their own will, where this implies an ability to intervene in the causal order and select from genuinely open options. But apart from the spooky (Strawson says, "panicky") metaphysics involved in this libertarian story, it just doesn't

[1] Strawson mentions two sorts of conditions that might cause us to withhold or moderate our reactive responses to perceived injury: exempting conditions and excusing conditions. Exempting conditions (which I discuss here) concern the moral capacity of the offending agent: is she someone who is genuinely fit to be held responsible? Excusing conditions concern whether or not a responsible agent (i.e., a non-exempted individual) is indeed responsible for a given act – perhaps she was coerced, or perhaps the act in question was an accident. I say a bit more about the difference between exempting and excusing conditions in the appendix to this paper.

serve any useful function. After all, if the story were true, our reactive attitudes would only be appropriately targeted by detecting when people possess and/or exercise their contra-causal free will. But how do we detect this? As everyone must agree, we can't do so directly; we have no "free-will-o-meters" to do this tricky job. So, this means we can only do so indirectly: that we must rely on some other property that attests (indirectly) to when someone is capable of exercising their contra-causal free will by way of attesting (more directly) to the appropriateness of holding them responsible. In other words, we need to be tracking some other property that attests (more directly) to whether or not they are fit to be held responsible. Hence, the capacity for exercising a libertarian free will is just an idle metaphysical wheel; it does no real work in underpinning our reactive practices.

That is Strawson's negative point. Now to the positive point. What makes people fit to be held responsible? What makes them an appropriate kind of target for our reactive attitudes? What property must they possess – and which we must be tracking – in order to make it appropriate for us to treat them as responsible agents?

First, we need to remind ourselves that, in Strawson's view, not all individuals are proper targets of the reactive attitudes. There are people who are cognitively and affectively abnormal in various ways (perhaps they are psychotic, deeply neurotic, or brain damaged in certain critical respects); and though these people may injure or even benefit us, we don't think there's any point in responding reactively to them (e.g., with resentment or gratitude, indignation or hurt feelings). Their handicap either makes them incapable of understanding the kind of demand expressed in our reactive attitudes (a demand, as Strawson says, for appropriate moral regard), or it makes them incapable of responding appropriately to that demand. They are unfit to be treated as "participants" in our shared moral practice, so it makes no sense to respond to them reactively.[2] Of course, we might

[2] In one sense, this paragraph may be highly misleading. As Strawson emphasizes, the distinction between those who are fit to be held responsible and those who are not is hardly black and white. This, indeed, is one important consequence of shifting from a metaphysical account of responsible agency, focussing on whether or not an agent possesses a libertarian free will, to a more naturalistic account of the sort Strawson favours. Responsible agency involves capacities

respond to them in all sorts of other ways: we may think it right to manage them, or restrain them, or provide them with some kind of treatment. And naturally this does not mean that they fall outside the scope of our moral regard. The point is just that we reserve our reactive responses for those whom we take to be capable of understanding what we are communicating through our reactive attitudes and are capable of responding appropriately.

Just what are we communicating though our reactive attitudes then? It is certainly part of our message that we expect, and indeed, demand, that individuals show one another an appropriate degree of moral regard. But given that our reactive attitudes are sensitive to judgments that we make about whether or not someone is a fitting recipient of these attitudes, the fact that we express them effectively communicates a good deal more. It says to the recipients that we don't despair of them as moral agents; that we don't view them "objectively" – i.e., as individuals to be managed or treated or somehow worked around; indeed, that we hold them accountable to an ideal of moral agency because we think them capable of living up to that ideal. So reactive attitudes communicate a positive message even in their most negative guise – even in the guise of anger, resentment, indignation. The fact that we express them says to the recipients that we see them as individuals who are capable of understanding and living up to the norms that make for moral community.

Our reactive attitudes will be well targeted, I have said, if the recipients can understand this message and have a capacity to respond in ways that show normative awareness of

---

that can be more or less well-developed, and developed in part (at least on Strawson's account) by how we engage with one another – i.e. to what degree and with what purpose we respond reactively to one another. Hence, I do not read Strawson as suggesting that our reactive responses are limited to those who are 'fully capable' moral agents, whatever that might mean. On the contrary: although he does not say much on this topic, it is a strength of his account that it allows for (limited) reactive responsiveness to those less able, precisely as a means of developing whatever capacity for responsible agency these individuals might possess (see, for instance, Strawson's discussion of young children, or therapeutic interactions of various sorts). Nevertheless, Strawson certainly thinks it possible that some individuals will be so disabled that ordinary, or even limited, reactive responsiveness is inappropriate. The point is to understand what sort of disability this might be – not from a biological perspective, but rather from a functional one. This is the question I am currently addressing.

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

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

Identifying this property puts the reactive attitudes themselves in a different light – and here I elaborate on Strawson's work with two further observations. First, there is a tendency (no doubt encouraged by the name Strawson gave them) to focus on the fact that reactive attitudes are backward-looking responses to the actions and attitudes of others. Yet, because they are themselves attitudes expressing the good or ill will of others, they will naturally prompt reactive responses in turn. After all, as Strawson points out, reactive responses reflect the fact that we care enormously about what attitudes others manifest towards us, and this will be true – perhaps even more true – when the attitudes in questions are *themselves* reactive attitudes: attitudes that in their nature have commented on the quality of our moral agency. So while reactive attitudes are backward-looking responses to the actions and attitudes of others, they have, more importantly, a forward-looking dimension, serving to elicit some further reactive response from the individuals to whom they're directed. It is this forward-looking dimension that is critical for understanding the power they have to scaffold the moral agency of others.

This leads to the second important observation. Reactive attitudes will function successfully in this scaffolding role, so far as they prompt normatively appropriate reactive responses in others. But since part of their aim is to elicit such responses when that aim is accomplished, these reactive attitudes are naturally answered and transformed, replaced by new reactive attitudes that are themselves appropriate responses to the reactive responses prompted by the original reactive attitudes. In other words, reactive attitudes perform their scaffolding role so far as they are normally embedded in dynamic trajectories of reactive exchange (see figure 1). These trajectories are actually what give the reactive attitudes that constitute them the meaning and power they have. Forgiveness is a good example. Forgiveness is a reactive attitude that serves to reaffirm the moral competence of the individual to whom it is directed. But it only makes sense as a reactive attitude – and only has the power it does – so far as it comes at the end of a trajectory of reactive exchanges occurring principally between a victim and a wrongdoer, but often involving the reactive responses of bystanders as well. Hence, if we want to

understand how reactive attitudes play a constructive role in making and sustaining moral community, we need to understand the trajectories of reactive exchange in which they are naturally embedded. (This point will become important in my final section.)

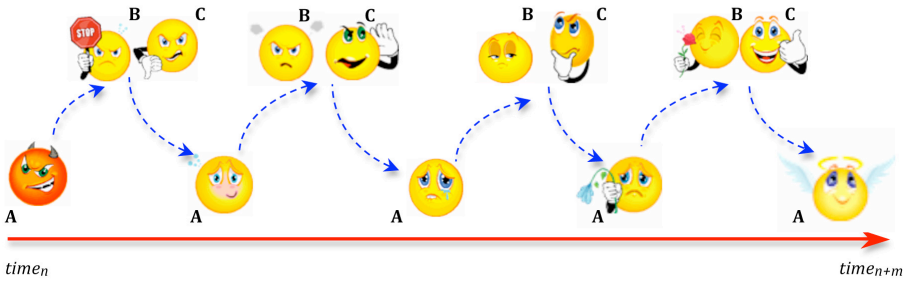

Figure 1 A sample trajectory of reactive exchange:
The forgiveness trajectory

So where have we got to so far? The philosophically important thesis that Strawson defends in his work is the metaphysical non-commitment thesis. According to this thesis, individuals must possess a certain property if they're to be appropriate targets of the reactive attitudes, if they are fit to be held responsible. However, this property has nothing to do with the metaphysics of human choice and so presupposes no commitment to the falsity of determinism (and no commitment to its truth either). It is simply the property of co-reactivity, or of being susceptible to the scaffolding dynamics of reactive exchange. To put this another way: individuals who are appropriate targets of reactive attitudes (i.e. who are fit to be held responsible) have a certain kind of character from which their actions flow – viz., the kind of character that is reactively responsive to, and so shaped under, the scaffolding influence of others' reactive attitudes.

I put the point this way for an important reason: A central plank of the libertarian metaphysician's argument is that our ordinary concept of responsibility is essentially connected to the following thought: responsible agents are agents that "could have done otherwise". They say that the only sensible interpretation we can give to this thought is metaphysical – i.e. it's a thought that only makes sense if we impute to agents special

(non-deterministic) causal powers. Call this the "causal reading" of "could have done otherwise". Yet if Strawson is right, there is another far more sensible interpretation we can and should give to this commonplace thought. Call it the "character reading" of "could have done otherwise". On this reading, the thought is that responsible agents have the kind of character such that they could have done otherwise, i.e., their character is not cut in stone, leading inevitably to the behaviour in question, but rather is "living and breathing", as George Eliot says, open to being formed and reformed under the regimen of reactive scaffolding. Thus, "could have done otherwise" expresses the commonplace view that, even though we certainly do not expect moral perfection from responsible agents on every occasion, we do expect them to have an ever-present moral sensitivity, where sensitivity is "operationalized" in terms of a disposition to respond appropriately – i.e. co-reactively -- to reactive attitudes.[3]

---

[3]It is interesting to note that, even from a very young age, children seem to divide the world into agents and non-agents, where (unconstrained) agents are viewed as being the kind of entity that 'could have done otherwise' (Nichols, 2004). In one study, Nichols presented 5 year olds with two types of scenarios, one involving people and the other involving inanimate objects. Sample scenarios were as follows (Nichols, 2004, p. 487):

> *Scenario 1*: Mary is in a grocery store and wants a candy bar. She chooses to steal the candy bar.
> *Test question:* Okay, now imagine that all of that was exactly the same and that what Mary wanted was exactly the same. If everything in the world was the same right up until she chose to steal, did Mary have to choose to steal?
> *Scenario 2*: A pot of water is put on the stove and heated up. The water boils.
> *Test question*: Okay, now imagine that all of this was exactly the same. If everything in the world was the same right up until the water boiled, did the water have to boil?

Nichols reports that participants were more likely to say that the outcome 'had to happen' in the physical scenarios than in the agential ones (though it's interesting to note that some children said the outcome had to happen even in the agential cases). What does this show? Nichols interprets these results as supporting the hypothesis that even young children are inclined to a libertarian view of agent-causation: that agents have a special kind of *causal power* that would have enabled them to choose other than what they chose to do. But this interpretation is certainly disputable, as Nichols himself point out. In particular, I suggest that the character reading of could have done otherwise is not ruled out by these results. On this reading, children duly note that agents *are* different from non-agents in so far as their actions flow from the type of person they are (their character), which character is manifested in the choices they make. Characters (persons) are the sorts of thing that develop and change. Hence, Mary (in the above scenario) 'could have done otherwise' – she is that kind of entity. By contrast, non-agents make no choices; they have no character; and have no potential to develop and change. Their behaviour is entirely determined by external conditions; hence, the outcomes described in the scenario "had to happen". In other words, on this hypothesis, when children hear the agent-involving scenario, they are more likely to pay attention to the fact that Mary is the type of entity that 'chooses', as against what the possible determinants of her choice might be on any given occasion. In short, the deterministic

The last point I want to make in this section is the following. While the metaphysical non-commitment thesis constitutes the philosophical core of Strawson's work, there is a second pragmatic thesis that I think is suggested by his view. I'll call it the *metaphysical corruption thesis*.

The concepts and practices of holding responsible embodied in our reactive attitudes make no metaphysical presuppositions. But, according to the metaphysical corruption thesis, these attitudes and practices may be sensitive to people's mistaken beliefs about the relevance of metaphysical views to their coherence and sustainability. That is to say, if people become convinced that such concepts and practices rest on embracing a libertarian conception of free will, and if they also accept the truth of determinism, then this can have a negative impact on the concepts and practices themselves. Hence, according to the metaphysical corruption thesis, there are specific practical dangers that flow from failing to embrace the metaphysical non-commitment thesis.

These practical dangers might arise at two levels. At the first level, the threat is quite direct. The worry is that if ordinary folk believe that the concepts and practices of holding responsible only make sense if people have libertarian free will, and if they become convinced that free will in this sense is an illusion, then they might lose faith in the idea that there is any real distinction to be made between individuals who are appropriate targets of reactive attitudes and those who are not, leading to substantial changes in the ways we conduct our inter-personal affairs. For reasons I won't go into here, Strawson was not particularly concerned with this possibility, and I think rightly so. (I discuss this issue further in the appendix to this paper, where I also consider Strawson's view in relation to some recent empirical studies conducted by Nichols and Knobe (2007).)

---

set-up is simply irrelevant to the kind of response they give – just as Strawson might have predicted. It will be interesting to see whether future empirical studies can rule this hypothesis out.

But there is another kind of threat operating at a higher-order level of social policy and institutional design. The Strawsonian concern could be put like this: Suppose that we – now as theorists of reactive practice – form the mistaken belief that such attitudes embody a concept of responsibility that depends for its coherence on a libertarian conception of free will. Then we, as theorists, are likely to misunderstand the internal dynamics of ordinary reactive practices and how they function to make and sustain moral community. But if we, as theorists, misunderstand this, then we'll have little chance of designing institutions that capitalize on, and so enhance, the best features of these practices; in fact, we may end up designing institutions that distort and disfigure them, to everyone's loss.

Although Strawson may not have paid too much mind to this possibility, I think the concern is real, having significant practical import in the field of corrective justice. I come to this issue in the third and final section of my paper. In the next section, I prepare the ground by briefly reviewing some recent work at the confluence of philosophy, cognitive psychology, evolutionary theory and the law that raises this concern most directly.

A new wave of moral-psychological research: a cautionary tale

In recent years, the study of moral psychology has become distinctly interdisciplinary. Philosophers, in particular, have been enjoined to get out of their armchairs and take account of findings in any of a number of related empirical disciplines: cognitive, social and evolutionary psychology, game theory, cognitive neuroscience, cognitive ethology, and the like. Some philosophers resist this trend on the grounds that empirical work is not really relevant to philosophical inquiry, conceived as a purely normative discipline. I am not of their number. In my view, there are many ways in which empirical work can aid and sharpen normative inquiry, keeping it tied to the world as we know it, even though it cannot resolve normative questions. Still, sympathetic as I am to this interdisciplinary trend, it can foster a tendency to shirk on the philosophical side of normative thinking, thereby leaving questionable assumptions unchallenged and

distorting the interpretation of empirical findings. In this section, I will examine some views that importantly contribute to current debates but that I think err in this direction.

The papers I discuss are both published in the *Philosophical Transactions of the Royal Society* in a special issue devoted to exploring the implications of findings in cognitive science for topics in the law. One is co-authored by Joshua Greene and Jonathan Cohen (2004); the other by Oliver Goodenough (2004); Goodenough's paper is, in fact, a direct response to Greene & Cohen's. These papers make radically different normative recommendations for systems of corrective justice, based on the authors' opposing views of how we should think of our reactive emotions. Nevertheless, despite the obvious ways in which these authors disagree, they embrace a common framework that utterly fails to take on board Strawson's central philosophical insight concerning the reactive emotions. My aim in this section is to show how these authors arrive at their opposing views within this common framework, while at the same time insisting that their views do not exhaust the field of possibilities. In fact, if what I've argued in the previous section is correct, the framework imposes a deeply impoverished understanding of the reactive attitudes that takes serious armchair reflection to counteract.

So what is the shape of this common framework? The authors mentioned above are heirs to a Strawsonian tradition in the following sense: they share the view that reactive emotions (or attitudes) are intimately tied to folk concepts and practices of responsibility. Hence, like Strawson, these authors are jointly concerned to focus theoretical attention on understanding how such reactive attitudes contribute to shaping our interpersonal lives. Moreover, since they embrace the thesis of determinism, they are jointly preoccupied with Strawson's issue; they too are concerned with what impact the truth of determinism might – and perhaps should – have on our reactive emotions and reactive practices. Here, however, these theorists depart radically from Strawson's own line, raising the spectre of the metaphysical corruption thesis. For they jointly reject, albeit without argument, Strawson's main philosophical insight: the metaphysical non-commitment thesis. That is, these authors simply assume that the folk concept of responsibility, as embodied in reactive emotions, is metaphysically committed to libertarian free will. And

this in turn raises the worry that such a mistaken theoretical belief can have damaging practical consequences, at least at the level of issuing policy recommendations. Indeed, in magnificent defiance of the metaphysical corruption thesis, these authors jointly assert that we can only design just and effective social institutions once we understand how a libertarian metaphysical commitment is woven into the very fabric of folk concepts and practices of responsibility. Needless to say, if they are not right about the metaphysical commitment, they are not likely to be right about how best to design just and effective social institutions. But I will save this discussion for the final section of this paper. My aim in what follows is just to present the views of the theorists in question, beginning with Greene and Cohen (2004).

The primary aim of Greene and Cohen's paper is to argue that discoveries in cognitive neuroscience have tremendous potential to effect reform in the law, especially with regard to our understanding of criminal responsibility and "just deserts". This may seem surprising, as they endorse the claim that the law makes no heavy metaphysical presuppositions about human agency in its assessment of criminal responsibility (see, for instance, Morse 2004). All that matters in the eyes of the law is that individuals have a general capacity for 'rational choice', which capacity is understood minimally as the capacity to act rationally in light of one's beliefs and desires. Hence, the legal determination of responsibility is in no way threatened by the thesis of determinism. However, while Greene & Cohen agree that the law "as written" is essentially compatible with the truth of determinism, they insist that this is simply not true of ordinary folk intuitions of responsibility, which are deeply libertarian. As they say, "In modern criminal law, there has been a long tense marriage of convenience between compatibilist legal principles and libertarian moral principles" (Green & Cohen, 2004, p. 1778). In their view, the reason the relationship has lasted so long is that ordinary folk have felt no pressure to engage in the kind of "esoteric theorizing" that tortuously reconciles current legal practices with a compatibilist doctrine of agential control. They simply assume that the law reflects their libertarian moral intuitions. However, Greene & Cohen predict that if push ever came to shove, and ordinary folk were forced to choose between rejecting current legal practices or accepting a compatibilist defense of them, they would reject

current legal practices, or at least the retributive elements of those practices. In their view, this is where contemporary neuroscience can play a major role -- by making push come to shove. As they say, "the legitimacy of the law depends on its adequately reflecting the moral intuitions and commitments of society. If neuroscience can change those intuitions, then neuroscience can change the law" (Greene & Cohen 2004, p. 1778).

So how is contemporary neuroscience to achieve this transformation? According to Greene and Cohen, the first thing it will do is to demonstrate beyond a shadow of a doubt that there is no "self" that is in charge of our actions behind all the neural firings that constitute brain activity. And if there is no self, there is no source of libertarian free will: what humans beings do just is a matter of how their neurons fire and how their neurons fire is completely determined by complex biological and environmental factors. As they say, "neuroscience can help people appreciate the mechanical nature of human action in a way that bypasses complicated arguments" (Greene & Cohen 2004, p. 1780). Once this happens, Greene and Cohen predict that we ordinary folk will experience conflict in our moral intuitions: this is because our ordinary notion of responsibility embodied in our reactive attitudes embeds a commitment to libertarian free will. And we will come to see this commitment as a hopeless error, hence, we will come to regard our own reactive attitudes as deeply misguided and unfair.

Will this change the shape of our everyday reactive practices? Alas, probably not, according to Greene and Cohen. They suggest that the reactive affective system is part of our evolved biological heritage: it is likely driven "by phylogenetically old mechanisms in the brain" and, hence, very unlikely to be cognitively penetrable (Greene & Cohen 2004, p. 1784). In other words, though we may come to regard our reactive attitudes (e.g., resentment, indignation, retributive anger) as embedding a hopeless error in the way we regard human agents, we may be stuck with such attitudes in the hurly-burly of everyday life.

In Greene and Cohen's view, this is very bad news. Unlike Strawson who, for quite different reasons, agrees that our everyday practices are no doubt immune to

metaphysical revision (see Appendix), Greene and Cohen regard such a change as normatively mandated. How then do we ordinary folk, handicapped by our evolutionary heritage, cope with this predicament? The way is not easy, in their view, but cognitively mediated processes may still come to our rescue. Although we may not be able to suppress our reactive attitudes in day-to-day life, we can at least "bracket" these attitudes (i.e., not be guided by them, and not cater to them) when it comes to designing and/or evaluating social institutions, especially those concerned with criminal justice. Indeed, since ought implies can, this is where our normative obligations must lie. I return to this point in the next section.

While Greene and Cohen represent one strand in this new wave of cognitive research, Oliver Goodenough represents a different strand, which nevertheless shares many of the same elements. As mentioned above, Goodenough (2004) acknowledges the basic Strawsonian point that reactive attitudes and practices embody our folk notion of responsibility. But, like Greene and Cohen, he rejects Strawson's metaphysical non-commitment thesis out of hand: that is to say, he does not argue the point, but simply accepts that the folk notion of responsibility presupposes a commitment to libertarian free will. Furthermore, since Goodenough accepts the truth of determinism, he agrees with Greene and Cohen that this folk commitment to libertarian free will is an error. He also agrees with Greene and Cohen that this error "may be deeply lodged in human cognitive and emotional psychology", hence will not be abandoned through deeper reflection on metaphysical issues (Goodenough, 2004, p. 1807). He thus arrives at Greene and Cohen's conclusion that good institutional design requires theorists to come to grips with how reactive practices depend on a libertarian metaphysics of free will – contrary to what the metaphysical corruption thesis explicitly warns against.

Now comes the interesting twist. Whereas Greene and Cohen see the commitment to libertarian free will as a hopeless error, showing the rot (as it were) at the heart of our reactive attitudes, Goodenough regards it as a useful fiction, playing a critical strategic role in regulating human interactions. Reactive attitudes, and more precisely punitive attitudes, with their inbuilt commitment to libertarian free will, have a strategic

evolutionary rationale. Moreover, this rationale not only explains why they are such a deep feature of human psychology; it also demonstrates, *contra* Greene and Cohen, why we would be foolish to bracket them in designing and/or evaluating our social institutions.

Here is a brief summary of Goodenough's (2004) argument explaining the evolutionary rationale for our commitment to the free will illusion (p. 1807): As psychologists have shown, punishment, especially by a third party (i.e., someone not directly involved in the offensive transaction), is effective at "stabilizing cooperative social structures" (Bendor & Swistak, 2001; Fehr & Fischbacher, 2004). However, to be effective, threats of punishment must be credible: they must involve a commitment to punish in the face of transgression (Dixit & Skeath 2004). Yet punishment is typically not without cost – think of how much we pay for prisons, protracted criminal trials, appeals, and so forth. Moreover, since there is often no direct material gain to punishers, especially third-party punishers, this cost must be borne "altruistically" (Fehr & Gachter 2002). Strategically, this means that punishers will not want to waste punishment on those for whom punishment has no effect – i.e., on those who truly cannot be changed or deterred through punishment. Yet this introduces an incentive for transgressors to make it seem to would-be punishers as if they could not be influenced through punishment, making those punishers less likely to inflict punishment on the transgressors. Now Goodenough asks: how could such feigned indifference to punishment be guarded against? One plausible evolutionary solution to this strategic problem is to build a "commitment" into human psychology: that is,,"design" human beings in such a way that they are cognitively programmed to see one another as having greater powers of rational agency and behavioural control than they actually have. Goodenough describes this default assumption as follows:

> The commitment is to treating the other agent as if he/she had the capacity to fully integrate the threat of punishment into its decision-making calculus, and to act accordingly, i.e., as if she/he had a kind of free will. Declaring this committed

> position both neutralises attempts at deception by the transgressor and to some degree forces the role of a considering agent on the other player
>
> (Goodenough 2004, p. 1807).

In other words, by grace of natural selection, we are committed to punishing would be transgressors because we inevitably default to viewing human beings in a certain light – viz., as self-directed agents possessed of a libertarian free will that gives them ultimate control of anything they do. Our reactive attitudes, especially our punitive attitudes, may be an ineradicable of human psychology, resting on a total fiction; but this is a good thing, according to Goodenough, something to be honoured in the context of social policy and institutional design. In short, theorists would do well to accommodate the following esoteric truth:

> However counter-factual the free will proposition may be in a deterministic world, it is a strategic fiction that underlies the productivity of a punishment rule…. Our free will intuitions may be false in the world of deterministic science and yet nonetheless effective in the world of strategic interaction [presumably the world in which we human beings, as social animals, have to survive. and indeed thrive]
>
> (Goodenbough, 2004, p. 1807)

The theorists whose work I have reviewed in this section come to radically opposed conclusions concerning the stance we should take towards our ordinary reactive attitudes: whether we should regard them, on the one hand, as embodying a hopeless atavistic error to be bracketed as much as possible in the context of social policy and institutional design; or, on the other hand, as embodying a strategically useful fiction that should be maintained as far as possible. This difference has profound relevance for their views on corrective justice, the topic to which I now turn in the final section of my paper. As we shall see, these authors embrace the standard opposing views in that domain: an enlightened welfarist approach focussed on deterrence and rehabilitation (traditionally associated with consequentialism), versus a strict retributive approach focussed on giving

offenders their "just deserts" (traditionally associated with deontology). However, my aim in the next section is to insist that these views are not exhaustive, but depend instead on endorsing a shared view of the reactive attitudes that we have good reasons to reject. Once that common assumption is rejected and we adopt a properly Strawsonian view of reactive attitudes, it is interesting to see what new theoretical terrain is made available for critical exploration.

Institutional significance of our theoretical commitments

Let me sum up the state of play so far. There's a big divide amongst theorists who maintain that our reactive attitudes express our ordinary intuitions/ judgments of responsibility and underpin our ordinary practices of holding responsible: On the one side, including Strawson himself, are those who think that these reactive attitudes and practices are metaphysically modest so far as they only presuppose co-reactivity on the part of responsible agents (the metaphysical non-commitment thesis). On the other side, representing a new research program in cognitive psychology (e.g., Goodenough and Greene & Cohen), are those who think that such attitudes and practices depend for their coherence on viewing responsible agents as possessed of a metaphysically expensive, libertarian (or contra-causal) free will. These two sides also differ on a related pragmatic issue: Strawsonians worry that theorists will do a poor job of institutional design if they fail to come to grips with the fact that ordinary concepts and practice of responsibility make no metaphysical commitments; they will fail to understand – and so properly exploit – the real internal dynamics of reactive attitudes and practices, thus undermining their power to make and sustain moral community (the metaphysical corruption thesis). The opposition, by contrast, thinks that theorists will do a poor job of institutional design if they fail to come to grips with the fact that reactive attitudes and practices depend on a libertarian metaphysics, whether this means compensating for a hopeless atavistic error or exploiting a strategically useful fiction.

In the first section of this paper, I argued in support of Strawson's metaphysical non-commitment thesis on the following grounds: Even if ordinary folk have a cognitive

tendency to buy into a metaphysical belief in libertarian free will at a quasi-reflective level, this belief cannot really be driving day-to-day judgments of responsibility as these are embodied in reactive attitudes and practices.[4] After all, even if individuals exercised this libertarian power of free will, there is no direct way of detecting when they have done so. So when we judge that people are responsible – and fit to be held responsible -- it must be on the basis of some other evidence – evidence, as I argued, for their being co-reactive agents: agents that are sensitive to the scaffolding dynamics of ordinary reactive attitudes. If this is right, we have good reason to reject the opposition's claims about how to conceptualize reactive attitudes when it comes to institutional design.

In this section, I want to explore why it matters. What is the practical significance of this debate for institutional design, specifically in relation to the issue of criminal justice? At this point, I can offer only a very brief sketch of the directions in which the different theories lead.

First of all, where would Greene and Cohen's hopeless-error view of reactive attitudes and practices lead? Greene and Cohen are explicitly reformist in their policy recommendations. They argue that, since our ordinary ideas of responsibility are deeply mistaken by presuming that agents have a power – libertarian free will – that they simply could not have, policy makers have an obligation to put these ideas aside in thinking about crime and punishment. But what does this leave? Appealing to a well-developed

[4] It seems clear from various studies (e.g., Nichols & Knobe 2007) that ordinary people (at least in North America, and I presume other OECD countries) have an overwhelming theoretical tendency to cash out the intuition that responsible agents 'could have done otherwise' in terms of their possessing something like a contra-causal free will. This is an interesting datum that deserves explanation. However, I stress that this is a *theoretical* tendency – meaning, it is the explanation that people are most naturally drawn to in giving a reflective account of what makes for responsible agency. Similarly, when people give a reflective account of the behaviour of physical objects (e.g. how an object will fall when thrown from a speeding train), they do so most naturally in Aristotelian terms. So here is my conjecture: folk theoretical proclivities do reveal something interesting about the way human beings are cognitively structured: certain theories for various natural phenomena are simply more intuitively appealing than others. However, the theories people are attracted to in their more reflective moments may have very little to do with what is guiding their behaviour in day-to-day life (including catching balls that are thrown to them, or discerning when someone is fit to be held responsible). This hypothesis is not ruled out by the Nichols and Knobe (2007) study, discussed at greater length in the appendix to this paper.

consequentialist line of argument (e.g., Smart, 1961), Greene and Cohen propose that the only just and equitable legal institutions are those that maximize overall social welfare. Of course, as they point out, there is nothing in the truth of determinism that suggests people cannot be conditioned by their external environment. Hence, if punishment, or the threat of punishment, serves to regulate behaviour in socially desirable ways, then, to that extent, it's all to the good. Furthermore, with this welfarist agenda in mind, incarceration in one form or another is a legitimate way to protect others in society, especially against those who cannot be deterred by the threat of punishment.

As advocates of "consequentialist legal reform", Greene and Cohen (2004) say they are promulgating a view that will "radically transform… our approach to criminal justice" (p. 1784). But how radical are their suggestions? For instance, in response to standard complaints levelled against this sort of view, they insist that their view justifies neither "extreme over-punishment" nor "extreme under-punishment" (p. 1783), presumably relative to current norms. Moreover, they argue that their view leaves in tact a number of distinctions currently recognized in the law, underpinning a variety of defences that diminish or undermine criminal responsibility (e.g. "duress", "diminished capacity", and so on). More importantly, they argue that their view leaves in tact the very notion of criminal responsibility officially recognized in the law, which – as Morse and others point out – already presupposes just the general and minimal capacity for rational agency that compatibilists favour. Given all this conservatism, one wonders just how "radically transforming" Greene and Cohen's consequentialist legal reforms would actually be.

Against such pessimism, Greene and Cohen remain convinced that there is much room for improvement. In their view, retributivism is the source of a number of ills in this domain: the idea that when we punish criminals, we are giving them their "just deserts" or "what they truly deserve" for the crimes they have (wilfully) committed. As they say, "our penal system is highly counter-productive from a consequentialist perspective, especially in the USA, and yet it remains in place because retributivist principles have a powerful moral and political appeal" (Green & Cohen, 2004, p. 1783). Of course, with regard to certain consequences, Greene and Cohen should insist that people's motivations

are not relevant when it comes to assessing the pros and cons of a given penal system. If retributive impulses deliver the appropriate amount of punishment from the perspective of maximizing overall social welfare, then those impulses may well be serving a useful societal function (as we shall see, this is Goodenough's view, discussed below). So, by "counter-productive", they could only mean that such retributive impulses do not deliver the best outcome from a social welfarist perspective after all.

Naturally, this is an empirical issue, and as such, I do not have much to say about it. But, surprisingly, nor do Greene and Cohen. They do suggest that, absent retributivism, people might be less keen on the death penalty (p. 1784) and perhaps that would conduce to overall social welfare (although they nowhere argue for this point explicitly). Another potential gain (which they also don't discuss) is that people might be more prone to support efforts at rehabilitating offenders, where presumably those efforts would be stripped of any taint of moral blame and focus instead on the kinds of conditions that debilitate individuals from acting in law-abiding ways (e.g., poverty, ignorance, social maladjustment, and so on) and doing what it can to address those conditions, both in the particular case and in society at large.

However, against these gains must surely be set an overall cost – one that Greene and Cohen do explicitly discuss, yet seem not to factor in to their consequentialist calculations. Recall that, in their view, retributive feelings (with their implicit commitment to free will) are likely to be an ineradicable feature of the human psyche, "driven by phylogenetically old mechanisms in the brain" (Greene & Cohen, 2004, p. 1784). Indeed, this is why Greene & Cohen think that the best we can hope to achieve is to bracket these feelings when it comes to policy making and institutional design. We will still experience these feelings in day-to-day life, but in "special situations" we can put them aside, adopting the more "detached" and "humane" perspective on criminal behaviour mandated by the truth of determinism (Greene & Cohen, 2004, p. 1784). Perhaps this dual perspective is genuinely possible for us – sufficient even for achieving the sort of reforms in our criminal justice system that Greene and Cohen would like to see. But at what psychic cost to individuals in society? Stuck as we are with our

atavistic tendencies to "see one another as free agents who deserve to be rewarded and punished for our past behaviours" (Greene & Cohen, 2004, p. 1784), how stable could our endorsement of these consequentialist reforms really be? We might continually remind ourselves of the neuroscientific findings that "graphically illustrate" how things "really are", but when faced with the next hideous crime in the morning news, our retributive impulses, in Greene and Cohen's own view, will come screaming to the fore. Moreover, just consider how much stronger such feelings would be if we, or our friends or loved ones, were the actually victims of the crime. Naturally, we need not – and often do not – act on all the feelings we experience. But to envision a situation in which we are so continuously at war with ourselves, with all of the likely social and political instability that portends, is surely a cost to be reckoned in any consequentialist assessment of the overall social welfare that would result from the reforms Greene and Cohen advocate.

In sum, Greene and Cohen's recommendations are both vague and problematic. On the one hand, they insist that once ordinary folk understand the importance of bracketing reactive attitudes (especially retributive impulses) in designing a genuinely fair and humane criminal justice system, they will naturally endorse a reformist agenda that is guided exclusively by consequentialist considerations of what produces the best societal outcome overall. On the other hand, even on Greene and Cohen's own consequentialist terms, it is unclear how radical such a reformist agenda would – or should – be. In the first place, empirical arguments are needed to support the claim that a punishment system designed on strict retributivist guidelines does not in fact deliver the best results from the point of view of deterring crime, keeping criminals off the street, and so on. And, secondly, even if those arguments were to fail, Greene and Cohen forget to take account of the psychic and, ultimately, social costs of designing a criminal justice system that runs counter to what in their view are ineradicable reactive attitudes that we will continue to experience in our day-to-day lives. In effect, what Greene and Cohen propose is a new and far less happy marriage of convenience between an intellectually sanctioned criminal justice system and persistent atavistic retributive feelings that will need to be continually policed and contained if the marriage is to survive. Perhaps in Greene and Cohen's utopia, there are no personal, social or political costs to this on-going domination of

detached objective reason over inter-personal affect, but that is rather a lot to hope for. A better empirical bet is that such envisioned reforms will come at considerable cost to overall social welfare, thereby undermining their consequentialist rationale. Hence, at the end of the day, it is not clear that their enlightened welfarist agenda, with its hopeless error view of reactive attitudes, has much to recommend it.

Oliver Goodenough is likewise sceptical of Greene and Cohen's reformist aspirations. Although he agrees with them that reactive attitudes embed a mistaken view of individual responsibility, he argues that consequentialists in particular should embrace the strategic advantages that flow from this illusion. In his estimation, evolutionary considerations and game-theoretic studies strongly suggest that the erroneous belief in libertarian free will ensures that punishment practices will be as effectively deterrent as possible, signalling to would-be transgressors that exempting conditions are hardly available, and thereby creating a big enough stick that most individuals would be hard-pressed to ignore. This argues for a criminal justice system that is organized around staunch retributive principles: holding people responsible and blaming them for their actions, even if, in some deeper causal sense, they are determined to do what they do. Goodenough agrees with Greene and Cohen that this is effectively how our current penal system is organized; so his policy recommendation is to leave well enough alone. On the strategic fiction view of reactive attitudes, a retributive approach to criminal justice is bound to deliver the best societal outcome overall.

What of Greene and Cohen's concern that a retributive system is hardly "fair" or "humane" in its treatment of individual offenders? Goodenough can certainly allow that committed consequentialists should be concerned with questions of individual fairness. However, it is hard to see how or why this should be a trumping concern. Deontologists may insist that it should be, but consequentialists are bound to reject such a move, keeping their eye on considerations of overall social welfare. Hence, if a concern with individual fairness is outweighed by other legitimate concerns, such as those connected with the psychic or social tolerability of various types of penal systems, then

consequentialists will just have to bite the bullet on questions of individual fairness. As Goodenough says:

> ....sadly, the efficacy of a punishment system may rest on a willingness to punish people who really could not help it. For better or worse, the Anglo-American approach to the law of responsibility…. is consistent with an in built commitment [to libertarian free will]
>
> (Goodenough, 2004, p. 1808).

The upshot of this new wave of cognitive research is that it does not take us very far in the field of criminal justice. In effect, we are left to choose between an old dichotomy of policy recommendations: on the one hand, an explicitly reformist, "enlightened welfarist" approach that refuses to endorse any deep notion of criminal responsibility and focuses instead on using the justice system to maximize deterrence and rehabilitation; and, on the other hand, a traditional retributive approach that embraces a deep notion of criminal responsibility, and insists on using the justice system, not primarily for deterrence, but rather to deliver merited punishment to individuals who wilfully engage in criminal behaviour. The only difference between these new wave discussions and the more traditional views is the change that is rung on the retributive approach. While retributivism is traditionally associated with the deontological concern of giving just deserts no matter what the consequences for overall social welfare – hence, with what deontologists view as trumping considerations of "individual fairness", Goodenough shows, perhaps surprisingly, that this approach may also recommend itself to welfare-maximizing, deterrence-oriented consequentialists. Indeed, if Goodenough is right, clear-eyed consequentialists *ought* to endorse retributivism, even if considerations of individual fairness (on their reckoning) speak against the approach rather than in favour of it. This is a bitter pill for consequentialists like Greene and Cohen to swallow. After all, their primary motivation is to advocate for a more humane and progressive criminal justice system. But now it seems as if they can maintain their consequentialism only at the cost of following in Goodenough's "enlightened retributive" footsteps, a dire option for those with a reformist agenda. Is there no way out of this dilemma?

An obvious suggestion is to give up on the consequentialism; but, in my view, this is the wrong course to follow. The root of these thinkers' problem is not in fact their consequentialism, but rather what Strawson himself identified as an "incomplete empiricism, a one-eyed utilitarianism" that distorts their consequentialist reasoning (Strawson 1974, p. 23). In the final telling paragraph of "Freedom and Resentment", Strawson writes:

> It is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes, in regulating behaviour in ways considered desirable; or to add that when certain of our beliefs about the efficacy of some of these practices turn out to be false, then we may have good reason for dropping or modifying those practices. What *is* wrong is to forget that these practices, and their reception, the reactions to them, really *are* expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. Our practices do not merely exploit our natures, they express them. Indeed the very understanding of the kind of efficacy these expressions of our attitudes have turns on our remembering this.
>
> (Strawson 1974, p. 25).

As I read Strawson, his main complaint against "one-eyed" consequentialists is their impoverished understanding of reactive attitudes and practices. In particular, because they are transfixed by the supposed metaphysical error they discern at the root of these attitudes and practices, they are concerned merely to bracket them or to exploit them in the crudest possible way (as devices calculatingly used for regulative purposes). Thus, they fail, each in their own way, to come to grips with how these attitudes and practices constitute a complex interconnected system of moral address. Indeed, as I argued in the first section, it is precisely because these attitudes and practices constitute a system of moral address that they are efficacious in scaffolding moral behaviour. To recap that argument: (1) certain normative demands are expressed in and through reactive attitudes, which demands are only properly directed towards agents that are sensitive to those demands and have a capacity to live up to them; (2) such sensitivity is shown through an

agent's co-reactivity: the disposition to respond reactively to others' reactive attitudes; (3) this disposition to co-reactivity means that reactive attitudes are naturally situated in dynamic trajectories of reactive exchange, where reactive attitudes serve to elicit some further response as much as they react to what has gone before; and, (4) while trajectories of reactive exchange can be more or less productive, they are psychologically and normatively most satisfying when they serve to restore, maintain or even generate a commitment to uphold the normative demands expressed in the reactive attitudes. Hence, well-directed and well-supported trajectories of reactive exchange serve the normatively critical function of scaffolding moral community. With this in mind, we can now turn to the criminal justice system, arguing that a Strawsonian approach to these matters would aim to make corrective practices co-reactive, helping to mimic in an institutional way the dynamics of co-reactive scaffolding.

From a theoretical perspective, it is important to see how this approach departs significantly from those discussed above. Consider, first, the differences with Greene and Cohen's "enlightened welfarism". A Strawsonian approach would not suggest that in punishment we simply look to the future good we can do by imposing deterrent penalties. It would vindicate a focus on the offender and on the wrong that was done. It would take the person as someone to be addressed in a properly reactive way, i.e. as a moral agent. It would not treat the person just as an instrument of social policy. To this extent it gibes more nearly with our intuitive understanding of what it means to treat an individual "fairly" (i.e., as deserving of an appropriate reactive response). Next consider the contrast with Goodenough's "enlightened retributivism": While endorsing the need to indict the offender in the manner I've just mentioned, a Strawsonian practice of corrective justice would embody a very different telos. Its aim would not be that of brutely imposing punishment simply to achieve maximum deterrent effect; rather it would strive to treat the offender as an appropriate target of moral address, holding up an ideal of responsible agency to the offender, while at the same time ascribing a capacity to live up to that ideal. In this way, it would treat offenders respectfully, recognizing their capacity for co-reactivity, and thereby scaffolding them in their efforts of restitution and reform. To this extent, it gibes more nearly with reformist ideals of what it means to treat

an individual "humanely" (i.e., as potentially always reclaimable in the context of moral community).

Among existing practices of corrective justice, is there any that would answer to this Strawsonian approach? Let me, in conclusion, mention one that looks to be on the right lines from the point of view of this paper. I cannot explore the approach here in any detail, but even a brief mention will help to support the claim that the considerations rehearsed here have a positive lesson in institutional design. The approach I have in mind is usually described as one of restorative justice. As an explicit movement it began from Mennonite-led initiatives in the 1970s. But one of the claims amongst supporters is that it represents a return to a type of community practice often found in "traditional" societies. Indeed, many restorative programs are being trialled in these societies -- e.g. amongst the first nations of Canada and the Maori of New Zealand; indeed, in New Zealand things have progressed to the point that these practices now dominate the juvenile justice system.

There is a great deal of variation in restorative justice programs, but certain features are particularly noteworthy. I bring these out by mentioning one of the more extensive empirical studies comparing such practices to more standard criminal justice procedures: the Reintergrative Shaming Experiment (RISE) conducted in Canberra, Australia from 1995-2000. RISE dealt with four different categories of offense: youth violent crimes (offenders under the age of 30); juvenile property offenses with personal victims (offenders under 18); juvenile shoplifting (offenders under 18); and drunk driving (offenders all ages). Here are the main elements in the restorative justice practice, as practiced in the Canberra RISE (discussed in Strang 2002; see also, Ahmed, Harris, Braithwaite and Braithwaite 2001):

i. Offenders admit guilt before agreeing to take part.
ii. They are then invited to participate in a conference to determine how their admitted offence should be rectified.

iii. The offender can bring a number of supporters (family members, friends) to the conference, as can the victim, assuming that the victim is willing to take part. Community representatives may also be present. The conference is chaired by a police officer.

iv. The formal purpose of the conference is to determine what the offender should do to make up for his or her offence but a crucial by-product is often that the offender comes to recognize the harm caused to the victim, and the victim comes to appreciate that recognition and, as often happens, the contrition that the offender displays.

v. Restorative justice supports the natural trajectory of reactive exchanges, discussed in the first section, and thereby provides the scaffolding that reactive attitudes can provide. This is good for the offender and good for the victim.

vi. In support of (v), the Canberra experiment revealed that restorative justice practices gave both greater offender and greater victim satisfaction than court proceedings -- victims in particular felt justice was done. In addition, there was some reduction in recidivism rates, more for those involved in violent crime.[5]

The elements of restorative justice as summarized here reflect central points that have been argued for in this paper. Under restorative justice initiatives, offenders are not treated in a crude consequentialist way as non-responsible targets of rehabilitation and/or deterrence; they are assumed to be fully responsible agents. Yet, in contrast with a traditional retributive approach, this presumption of responsibility is not focused simply on the crime, justifying a punitive response that ensures the offender receives appropriate payback for the wrong he or she has wilfully done. Instead, offenders are encouraged to take responsibility for their wrongdoing by coming to see themselves, not only as agents of crime, but more importantly as agents both of restitution and of recommitment to the standards of moral community. While such restitution may be legally required and

[5] For more detailed analysis of how recidivism rates appear to depend on the offense category (at least partially), see (Sherman, Strang and Woods 2000)

enforced, what cannot be required are the reactive emotions that victims, offenders and other stakeholders often feel in the context of a restorative justice conference as they try to come to grips with their own experiences of the crime and the meaning it should have in their shared community. And yet it is this reactive dynamic – in particular, specific trajectories of reactive exchange – that seems to correlate most nearly with genuine recommitment to the standards of moral community; at least as this recommitment is measured in terms of recidivism rates, as well as a general feeling amongst all the stakeholders that justice was served. But this fact is not well theorized or even recognized in the context of criminal justice. As John and Valerie Braithwaite observe:

> The genius of restorative circles is their collective emotional dynamics. At the moment, the research literature on restorative justice has not risen to the challenge of capturing these dynamics in research reports… The result of this failing is that even the most literate of criminologists and criminal lawyers understand restorative justice in terms of material reparation to victims, rather than in terms of symbolic reparation which all evidence to date suggests is more important.
>
> (Braithwaite & Braithwaite, 2001, 59).

This paper is the beginning of an attempt to fill that theoretical gap. It suggests that a broadly Strawsonian perspective on the importance of reactive dynamics for scaffolding moral agency and moral community can have significant institutional implications. It indicates why restorative justice holds out real promise as a just, humane and effective innovation in criminal justice. And it even gives some theoretical guidance to practitioners of various restorative justice initiatives, accounting for at least one factor that explains why some of these initiatives may be more successful than others. More work must be done to defend these claims, both empirical and philosophical. But I hope there is sufficient promise in what I have argued here to make that work worth undertaking.

Appendix

The metaphysical corruption thesis holds that mistaken beliefs about the relevance of metaphysical views to our ordinary concepts and practices of holding responsible can have a corrosive effect on those concepts and practices themselves. Philosophers familiar with Strawson's paper may find it surprising that I would find such a thesis compatible with his views on the following grounds: In "Freedom and Resentment", Strawson explicitly takes up this issue, asking both a predictive and a normative question:

> What effect would, or should, the acceptance of the truth of a general thesis of determinism have upon … [our] reactive attitudes? More specifically, would, or should, the acceptance of the truth of the thesis lead to the decay or the repudiation of all such attitudes? Would, or should, it mean the end of gratitude, resentment, and forgiveness; of all reciprocated adult loves; of all the essentially *personal* antagonisms?
>
> (Strawson 1974, p. 10)

Unsurprisingly, his response to the normative question is a decided "no". This just follows from his endorsement of the metaphysical non-commitment thesis: Since reactive attitudes and practices do not presuppose any metaphysical commitments – in particular, a commitment to libertarian free will – accepting the truth of determinism generates no rational obligation to abandon them.[6]

---

[6] Strawson's answer to the normative question is, in fact, more nuanced than I indicate here. Specifically, he argues, first, that the question is fundamentally absurd since there are no realistic conditions under which it could ever arise in such a comprehensive form. But, secondly, and more importantly, since such a choice would not mandated by the truth of determinism – again, this is the import of the metaphysical non-commitment thesis – then the only rational grounds on which we could ever abandon such concept and practices must hinge on "an assessment of the gains and loses to human life, its enrichment or impoverishment" – an assessment that is quite independent of accepting (or rejecting) the truth of determinism (Strawson 1974, p. 13).

More interesting, for present purposes, is Strawson's response to the predictive question. In a famous passage he remarks:

> The human commitment to participation in ordinary inter-personal relationships is, I think, too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might so change our world that, in it, there were no longer any such things as inter-personal relationships as we normally understand them; and being involved in inter-personal relationships as we normally understand them precisely is being exposed to the range of reactive attitudes and feelings that is in question.
>
> This, then, is a part of the reply to our question. A *sustained* objectivity of inter-personal attitude -- [i.e., of treating others as mere objects of manipulation or management], and the human isolation which that would entail, does not seem to be something of which human beings *would* be capable, even if some general truth were a theoretical ground for it
>
> (Strawson 1974, pp. 11-12; my emphasis).

Of course, given the metaphysical non-commitment thesis, Strawson doesn't think that the truth of determinism really provides any such theoretical ground. But here the question is: what if we thought that it did? And his answer seems to indicate that he endorses something closer to a metaphysical *non*-corruption thesis: our theoretical commitments just wouldn't make much of a difference to our reactive attitudes and practices, no matter how relevant, in more reflective moments, we might (mistakenly) take such commitments to be.

Interestingly, recent empirical work seems to support this Strawsonian prediction. Nichols and Knobe (2007) designed a study to test how people's judgments about moral responsibility might vary under different conditions. Specifically, their studies indicate that: (1) under abstract conditions, an overwhelming majority of people (86 per cent) are inclined to judge that "a person" cannot be fully morally responsible for their actions in a

deterministic universe;[7] but (2) under more concrete affect-inducing conditions, where a specific protagonist does something morally wrong (e.g., "Bill stabs his wife and children"), people are much more inclined to judge that the protagonist can be fully morally responsible, even in a deterministic universe. Indeed, in light of a follow-up study, Nichols and Knobe conclude that the more people's reactive emotions are stimulated by a scenario, the more prepared they are to view the protagonist as fully morally responsible, even in a deterministic universe.[8]

Although I can't here discuss their work in much detail, it's worth pointing out that Nichols and Knobe take these studies to support something like a dual process account of how people make judgments about moral responsibility: Under emotionally neutral conditions, people's judgments are subserved by a "more abstract, theoretical sort of cognition", guided by a particular quasi-reflective understanding of how moral responsibility is linked to the metaphysics of human choice (and here it appears that ordinary folk tend to be overwhelmingly incompatibilist in their intuitions).[9] But, under conditions that trigger reactive emotions, people are more likely to make judgments about moral responsibility that are in line with compatibilism (i.e. that are unaffected by the thesis of determinism). Nichols and Knobe suggest that such "compatibilist" judgments are generated by affect-involving processes that constitute a quite distinct

---

[7] The exact form of their question was: 'in [deterministic] Universe A, is it possible for a person to be fully morally responsible for their actions?'

[8] The follow-up study involved presenting participants with a 'high-affect' and 'low-affect' scenario, each of which was presented in a deterministic universe and a non-deterministic universe with respect to human choice. In the high-affect condition, participants were presented with the following question: "As he has done many time in the past, Bill stalks and rapes a stranger. Is it possible that Bill is fully morally responsible for raping the stranger". In the low-affect conditions, participants were presented with the following question: "As he has done many times in the past, Mark arranges to cheat on his taxes. Is it possible that Mark is fully morally responsible for cheating on his taxes?" Findings were as follows: Assuming a deterministic universe, 64 per cent of participants said Bill could be fully morally responsible, whereas only 23 per cent of participants said that Mark could be fully morally responsible. In the non-deterministic universe, 95 per cent thought Bill could be fully morally responsible, and 89 per cent thought Mark could be (Nichols and Knobe 2007, pp. 675-677)

[9] I say 'quasi-reflective understanding' because I don't mean to suggest that this view is arrived at through any deep reflection; only that it represents the natural folk way of theorizing about what makes us responsible agents.

psychological subsystem.[10] Since this subsystem is presumed to be relatively impenetrable to reflective cognition, any consciously held beliefs (e.g., that the universe is deterministic, and that full moral responsibility is not possible in such a universe) would have little impact on its workings. In other words, though they don't use this word, Nichols and Knobe would explain Strawson's metaphysical non-corruption thesis by appeal to the "modular" nature of an affect-involving psychological subsystem that willy-nilly generates judgments of moral responsibility under concrete, affect-inducing conditions.[11]

While the current trend in cognitive psychology is towards such dual process models, the Nichols and Knobe account does not do justice to the point Strawson is trying to make. The key idea behind Nichols and Knobe's proposed model is that our affective reactions are driving (in a causal sense) our assessments of moral responsibility. But, in

[10] Nichols and Knobe describe this subsystem as generating 'compatibilist intuitions'. I presume all they mean by this is: 'intuitions or judgments in line with a compatibilist theory of moral responsibility'. After all, if they're right to posit such a subsystem, the 'implicit' theory of moral responsibility according to which this subsystem operates has yet to be determined. For all their studies show, the implicit theory might be libertarian, with agential action automatically coded as produced by a *sui generis* act of will (no matter what views of agential action may be held at the level of conscious belief). This seems to be Greene & Cohen's (2004) view, discussed in Section 2. Alternatively, the implicit theory might make no such metaphysical presuppositions, simply coding agential action as action produced by appropriate psychological antecedents (beliefs, desires, intentions), which psychological states are themselves sufficient to trigger an affective response. Perhaps this is Nichols and Knobe's own view, although they do not discuss this issue explicitly.

[11] In a more puzzling part of their paper, Nichols & Knobe raise the question of whether such affective processes should properly be viewed as delivering 'more reliable' judgments about moral responsibility than the consciously endorsed theory (an 'affect competence model'), or whether the consciously endorsed theory should be viewed as delivering more reliable judgments about moral responsibility, with affective processes skewing these judgments whenever they are brought into play (a 'performance error model' of affective processes). I say this is more puzzling, because it's not quite clear what they mean by 'more reliable' in this context: more reliable, in the sense that the judgments so delivered better conform to people's (possibly benighted) underlying ideas about moral responsibility; or more reliable, in the sense that the judgments so delivered conform to the true view of moral responsibility? I assume Nichols & Knobe cannot mean the latter – otherwise their stated reason for preferring the 'performance error' model don't make much sense (viz., that it better captures the patterns of judgment observed in participant responses). However, if Nichols & Knobe mean the former, the whole question of 'reliability' seems off base, since their data may be indicating that people do not have any consistent underlying view of moral responsibility for such judgments to track (reliably or unreliably).

Strawson's view, this gets the causal order the wrong way around: Reactive emotions do not *cause* our judgments of moral responsibility; they are *expressions* of those judgments in particular concrete situations – namely, situations that reveal the quality of a person's good or ill will towards us, or towards other individuals. Hence, they are expressions of a basic moral stance we take towards others: Namely, a stance that assumes they are fit to be held responsible. Of course, this implies that reactive emotions will be sensitive to considerations that bear on whether or not we think it appropriate to treat others as responsible for what they do. And this is just what we find: If we come to realise that it is inappropriate to blame someone for a harm they did, then we cease to be resentful. Strawson emphasizes this point in arguing that "excusing" and "exempting" considerations trigger an end to resentment and the like. Excusing considerations go to the question of whether a responsible agent bears responsibility for a particular act (maybe she was coerced or did what she did accidentally), whereas exempting considerations go to the question of whether an agent is indeed a responsible agent – i.e. fit to be held responsible (maybe she suffers from a serious psychological disorder). Strawson's point is that, in the context of our everyday interactions, our reactive attitudes are deeply (though, perhaps, imperfectly) sensitive to a wide variety of such excusing and exempting considerations. With respect to exempting considerations, he simply adds that these will have nothing to do with abstract and perfectly general metaphysical beliefs, but only with more pedestrian and individually specific signs of moral incompetence.

Does this alternative Strawsonian reading of reactive attitudes, and the considerations to which they're sensitive, provide an alternative explanation of the Nichols and Knobe experimental results? Well, certainly the Strawsonian interpretation is not ruled out. The toy concrete scenarios, unlike the more abstract question, do at least provide a little snapshot of real inter-personal engagement; hence, it is no surprise that respondents are more prepared to fall back on their ordinary ways of assessing moral responsibility in these contexts, as opposed to being guided by abstract (and likely ill-understood) metaphysical doctrine. Of course, the toy scenarios provide no information about the protagonists' capacity to operate as morally competent agents in Strawson's sense. Still, the default assumption of ordinary interpersonal interaction is that such competence

exists unless and until proven otherwise (this is the force of Strawson's claim that exempting conditions are not the norm). Hence, it's no surprise that, in the absence of a more specific indications of moral incompetence, respondents will view the protagonists as responsible agents (by responding to them reactively), especially when the stakes are high – especially when the protagonists have manifested the kind of significant disrespect or ill-will towards particular others that would normally demand a moral response from those in the surrounding community.[12] In effect, such high-stakes situations make it morally problematic for putatively disinterested bystanders (e.g., respondents to questionnaires) to stand idly by, not even offering so much as a breath of moral condemnation just because exemption is claimed for the protagonist on abstract and unfamiliar grounds.

In light of these observations, let's now return to Strawson's own explanation of his metaphysical non-corruption thesis. The reason reactive attitudes are relatively immune – in everyday contexts -- to abstract theoretical considerations about determinism and human freedom is that these attitudes express our expectations of, and respect for, one another as morally responsive and responsible agents; hence, reactive attitudes presuppose a view of others on which many of our ordinary inter-personal relationships depend. To suppress or distance ourselves from these attitudes is to take a view of others – an objective view -- that is deeply inimical to this inter-personal view. As Strawson says:

> … [These views] are not altogether *exclusive* of each other; but they are, profoundly, *opposed* to each other. To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided… The objective attitude may be emotionally toned in many ways, but not in all ways: it may include repulsion or

[12] In this context, it's worth noting that someone's cheating on their taxes does not demand the same kind of moral redress from morally responsible bystanders. No ill will has been manifested, or harm done, to particular identifiable others. So respondents can afford to relax their moral vigilance, and let more idle metaphysical speculations weigh in on (relatively inconsequential) responsibility judgments.

> fear, it may include pity or even love, though not all kinds of love. But it cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships; it cannot include resentment, gratitude, forgiveness, anger, or the sort of love that two adults can sometimes be said to feel reciprocally for each other"
>
> (Strawson, 1974, p. 9)

I quote at length to emphasize that, for Strawson, what differentiates these two views is not the level of affect we experience towards others: we might be deeply afraid of someone towards whom we think it right or appropriate to take an objective attitude (e.g. the murderous psychopath running loose at night in our city). Rather what differentiates these two views is the stance we take towards others, which stance will determine the kind of affect it's possible to experience in relation to them. That is, do we treat them as appropriate subjects for moral address (exposing them to the range of our reactive attitudes)? Or do we treat them as individuals to be managed, whether for our good, for society's good, or even for their own good?

Both stances are available to us. Indeed, as Strawson says again and again, the objective stance is one we *ought* to adopt towards those who are not fit to be held responsible, and adopted to varying degrees depending on moral capacity. Furthermore, it's a stance we can take towards others for various reasons having nothing to do with the practical assessment of another's moral competence. As Strawson () points out, "we *have* this resource and can sometimes use it: as a refuge, say, from the strains of involvement; or as an aid to policy; or simply out of intellectual curiosity" (pp. 9-10). But (and this is the essence of Strawson's metaphysical non-corruption thesis) resorting to the objective stance can hardly be the norm, since it precludes the kinds of inter-personal relationships that are absolutely central to our human way of life. Thus, practically speaking, the option of systematically suppressing or distancing ourselves from reactive attitudes is simply not available to us, no matter what we might come to believe about the propriety of our moral concepts in more benighted philosophical moments.

Now we come to the nub of the point I want to make in this Appendix. It's clear that Strawson endorses a metaphysical non-corruption thesis at the practical level, at the level of day-to-day human interactions. But his thesis has nothing to do with the cognitive impenetrability of affective psychological processes, as some cognitive scientists might be tempted to suppose. Rather, his thesis depends on the fact that there are certain kinds of inter-personal relationships we cannot do without if we're to live a recognizably human form of life. Thus, his thesis is perfectly consistent with the idea of our removing ourselves from these relationships to some degree, some of the time. And in fact, we may think it rationally desirable to do so for certain constrained purposes; Strawson explicitly mentions assuming the objective attitude as an aid to social policy. But however valuable this resource may be, it is clearly a double-edged sword. In particular, from the objective stance, it's easy to lose sight of -- or at any rate discount – certain key features of our ordinary concepts and practices of holding responsible, leading to their systematic mischaracterization.

Why is this important? Because it suggests that there is another, higher-order level at which our ordinary concepts and practices of responsibility, as embodied in reactive attitudes, can be threatened by failing to grasp the import of Strawson's metaphysical non-commitment thesis. This is what I am calling the *metaphysical corruption thesis*. Once again, it goes like this. Suppose we, in our cooler theoretical moments, come mistakenly to believe that our ordinary concepts and practices of responsibility depend for their coherence on a libertarian conception of free will. Then we, as theorists, are likely to misunderstand the internal dynamics of ordinary reactive practices and how they function to make and sustain moral community. But if we, as theorists, misunderstand this, then we'll have little chance of designing institutions that are well suited to actual reactive practices. This paper has been an attempt to show why this concern is real.

References

Ahmed, E., Harris, N., Braithwaite, J., & Braithwaite, V. (2001). *Shame Management Through Reintegration*. Cambridge, UK: Cambridge University Press.

Bendor, J. & Swistak, P. (2001). The evolution of norms. *The American Journal of Sociology*, *106*, 1493-1545.

Braithwaite, J. & Braithwaite, V. (2001). Shame, shame management and regulation. In E. Ahmed, N. Harris, J. Braithwaite and V. Braithwaite (Eds.), *Shame management through reintegration*. Cambridge, UK: Cambridge University Press, 3-70.

Dixit, A., & Skeath, S. (2004). *Games of strategy*. New York: W.W.Norton.

Fehr, E. & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63-87.

Fehr, E. & Gachter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137-140.

Goodenough, O. R. (2004). Responsibility and punishment: whose mind? A response. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*, 1805-1809.

Greene, J. & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, *359*, 1775-1785.

Morse, S. J. (2004). New neuroscience, old problems. In, B. Garland (Ed.), *Neuroscience and the law: brain, mind, and the scales of justice* (pp. 157-198). New York: Dana Press.

Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind and Language, 19,* 473-502.

Nichols, S. & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, *41*, 663-685.

Sherman, L. W., Strang, H., & Woods, D. J. (2000). *Recidivism Patterns in the Canberra Reintegrative Shaming Experiment (RISE)*. Canberra: Australian National University.

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

Strang, H. (2002). *Repair or Revenge: Victims And Restorative Justice*. Oxford, UK: Clarendon Press.

Strawson, P. (1974). Freedom and resentment. *Freedom and Resentment and Other Essays* (pp. 1-25). London: Methuen.