The Nature of Reactive Practices: Exploring Strawson's Expressivism[1]

Thaddeus Metz

Philosophy Department
University of the Witwatersrand, Johannesburg
Private Bag 3
WITS 2050
SOUTH AFRICA
Email: Thaddeus.Metz@wits.ac.za

Abstract:
I aim to answer the questions of whether reactive practices such as gratitude and punishment are inherently expressive, and, if so, in what respect. I distinguish seven ways in which one might plausibly characterise reactive practices as essentially expressive in nature, and organise them so that they progress in a dialectical order, from weakest to strongest. I then critically discuss objections that apply to the strongest conception, questioning whether it coheres with standard retributive understandings of why, when and where the reactive practice of punishment is justified.

1. Introduction

It is amazing how P. F. Strawson's 'Freedom and Resentment' (1962[2]) continues to prompt reflection. Upon my most recent reading, I was struck by the way in which Strawson characterises reactive practices as expressive, and apparently essentially so. Reactive practices are, roughly, responses to others' behaviour that are made for no result metaphysically distinct from the responses themselves and that are regulated by judgments of degrees of responsibility. Key examples are punishment based on the finding that someone has done a wrong of a certain nature (and not, say, the expectation of deterrence) and gratitude based on the acknowledgement that someone has gone out of her way for you (and not, e.g., the aim of receiving more benefits for yourself as a result of the other appreciating your gratitude). Strawson is naturally read as maintaining that reactive practices such as these are partially *constituted* by the 'reflection' (63), 'display' (67), or 'manifestation' (80) of attitudes, particularly those of good- or ill-will.

Call this characterisation of the nature of reactive practices 'expressivism'. In this article, I aim to answer questions such as: What is the most plausible construal of expressivism? In what sense is it most likely to be true that, say, reactive punishment and gratitude are essentially expressions of certain attitudes? Who must do the ex-

1 For valuable input, I am grateful to participants in the Conference in Memory of P. F. Strawson that was organised by the University of the Witwatersrand Philosophy Department, and, especially, to David Martens.

2 All parenthetical page references in this article refer to this text.

pressing, what must be expressed, how must it be done, and to whom? Is expressivism in fact true? Is it part of the essence of reactive practices to reveal attitudes, or can a reactive practice exist that does not perform this function?

I find such questions interesting in their own right as a matter of social philosophy. In addition, the answers may have an important bearing on the ability to obtain coherence in (deontological) moral thinking. For instance, as I discuss below, it is not clear that any form of expressivism coheres with retributive intuitions about the proper function of legal punishment and about the universal scope of its justification.

I begin by analysing the concept of reactive practices (or defining what the phrase 'reactive practices' means), in order to clarify the behaviour that I seek to apprehend (section 2). Then, I distinguish seven distinct ways in which one might plausibly characterise reactive practices as inherently expressive, and argue that one specification is more plausible than other, more common and intuitive ones (section 3). More specifically, I organise these seven principles in terms of a developmental logic, starting with the weakest version, making objections to it, presenting a new principle that avoids the objections, making new objections, proffering yet another principle that avoids all previous objections, and so on, until I reach the most defensible instance of expressivism. In the following section, I present two objections to this best form of expressivism that question whether such a conception of reactive practices as applied to punishment fits with retributive theories (section 4). For now, I find myself unable to judge whether the objections are successful or not. So, while I specify which form of expressivism is most likely to be true, I do not make any definitive claim as to whether it is in fact true or not. I briefly conclude the paper by summarising what has been accomplished and noting issues to explore in the future in order to draw a firm conclusion about the truth of expressivism (section 5).

2. The Concept of Reactive Practices

My aim in this article is to ascertain what reactive practices are essentially, with expressivism and non-expressivism being competing conceptions of its inherent nature. For these controversial conceptions to be genuine rivals, they must concern the same thing, that is, they must share a common, uncontroversial concept of reactive practices about which they further disagree. In this section, I articulate this uncontested concept of reactive practices, the definitional core that makes it the case that a conception is one *of reactive practices* as opposed to something else. In short, what is the behaviour that Strawson maintains is essentially expressive and that I am not so sure is in fact expressive? Several examples of this behaviour include: criticising, blaming, punishing, praising, rewarding and acting gratefully. What do all these actions have in common that expressivists and non-expressivists can agree on as a matter of stipulation?

First off, reactive practices are actions. Strawson often speaks of reactive 'attitudes' or 'feelings', but mental states such as intention, conation, affection and emotion must be distinguished from volition, the behaviour at the core of a reactive *practice*.

In addition, reactive practices are not merely actions, but reactions. They are responses to 'the quality of….wills' (70), i.e., to the choices people make. So, helping someone merely because she is a person would not count as a reactive practice, but helping her because she has helped you would.

These reactions can be self- or other-regarding responses, i.e., directed toward oneself or toward others. Strawson usually cashes out reactive practices in terms of interpersonal behaviour, viz., actions that one could not perform if one were alone, such as

punishing another. However, 'self-reactive' (71) practices exist, say, acquiescing to punishment by another or intentionally inflicting harm on ('punishing') oneself.

In addition, the reactions can be personal or impersonal, i.e., a response either to actions done to oneself, which are personal, or to actions done to others, which are impersonal. Criticising someone because she has wronged you would be a personal reactive practice, while criticising her because she wronged another would be an impersonal one.

The self/other and personal/impersonal distinctions crosscut each other, meaning that there are four logically possible types of reactive practices. An example of a self-regarding, personal reactive practice would be rewarding yourself for a job you did well; an other-regarding, personal reactive practice would be blaming another person for having harmed you; a self-regarding, impersonal reactive practice would be rewarding yourself for having sacrificed on behalf of someone else; and an other-regarding, impersonal reactive practice would be blaming another person for having harmed someone other than you. This typology of reactive practices can be considered common ground among expressivists and their critics; I see no reason for either side to question this.

Furthermore, it is uncontroversial that reactive practices have *some* kind of association with reactive attitudes. For instance, punishment is usually associated with resentment and censure, while gratitude has some kind of relationship with appreciation and good-will. Even non-expressivists can accept that there is often some kind of contiguous or causal relation between reactive practices and reactive attitudes. What they question is Strawson's suggestion that reactive practices are *essentially manifestations* of reactive attitudes.

Yet another uncontested feature of reactive practices is that they are always associated with what Strawson calls a 'participant' attitude or orientation in contrast to an 'objective' one (66). A participant orientation is non-consequentialist and 'retributive' in the broad sense of seeking 'pay back' and no result beyond the reaction itself. The agent's reason for reacting to someone's action is not the expectation of a certain desirable consequence in the future. In other words, a reactive practice is analytically an action for which the motivation is either 'present-' or 'backward-looking', as opposed to the objective disposition's 'forward-looking' rationale. Punishing for the sake of incapacitation, deterrence or moral reform would be instances of an objective attitude, for these are all potential consequences of the punishment. In contrast, punishing in order to give someone what he deserves merely for having committed a crime would be an instance of a participant attitude.

The last major feature of a reactive practice that both expressivists and their critics can accept is that they are regulated by conditions of mitigating factors. Reactive practices are reactions that are contingent on, and the intensity of which is determined by, judgments of the degree to which the agent was competent in general and responsible for the given act in particular. For instance, it would be inconsistent with a reactive practice to punish someone for performing an act it is known that he could not avoid by virtue of insanity, or to act gratefully toward someone for an action that benefited you but that it is known that she did not mean to perform. More specifically, given the participant orientation, it is part of a reactive practice that you would not punish the insane individual *because* he was insane when performing the act (not because you expect bad consequences to result from his punishment), and it is part of a reactive prac-

tice that you would not be grateful to someone who did not intend to perform an action that helped you *for the reason that* she did not so intend.

Putting these pieces together, here is a rough, summarising definition of a 'reactive practice': an act that is a response to an action done by oneself or others with regard to oneself or others, which response is for the sake of no result distinct from the response itself and is governed by judgments of excusing conditions. Strawson apparently believes that there is an additional feature inherent to a reactive practice, namely, that it is expressive of attitudes such as good- and ill-will.[3] It is this claim that I critically investigate in the rest of this article.

3. Expressive Conceptions of Reactive Practices

In 'Freedom and Resentment', Strawson repeatedly contrasts the instrumental function of the objective attitude with the expressive function of a reactive practice. According to him, an objective orientation is concerned with bringing about valuable states of affairs, whereas the non-instrumental orientation that constitutes a reactive practice is a matter of expressing certain attitudes, usually of good- or ill-will. Here is a representative passage from Strawson:

> 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….What *is* wrong is to forget that these practices, and their receptions, 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 (80).

Although Strawson does not unambiguously say that reactive practices *essentially* or *inherently* express attitudes, the above passage may be fairly read that way; doing so conflicts with nothing else he says, and it raises important philosophical issues. In this section, my aims are to differentiate seven ways in which reactive practices might be constituted by expressions of attitudes and to argue that one of them is more plausible than the others. I start with the weakest version of expressivism and end with the strongest.

> E1: A reactive practice is essentially (in part) a communication of one's occurrent attitude about a person's action that is consequent to the action and is directed to the one who has acted.

This principle is motivated by the most common instances of punishment and gratitude *qua* reactive practices. In the normal case of punishment, an offender stands before a judge, who, in pronouncing and enforcing a sentence upon the offender, thereby conveys his disapproval of the offender's wrongful behaviour. And when we typically encounter gratitude, it takes the form of one who has benefited saying 'thank you' to her benefactor. These everyday occurrences underwrite E1, which basically says that punishment and gratitude as reactive practices are ways for one agent to communicate something about another agent to that agent. By 'communication' here, I have in mind a Gricean notion of, very roughly, one agent intending to use a symbol to convey a

3 And not merely of a participant attitude.

mental state to someone who in fact becomes aware of this mental state by virtue of the use of the symbol.[4]

However, there are counterexamples to E1, i.e., cases in which punishment and gratitude intuitively take the form of reactive practices and yet do not include the relevant sort of communication. For instance, suppose I feel grateful to a benefactor who is now dead. I cannot help him, but, since I can help his daughter, I can act gratefully with respect to him by helping her, and without the expectation of long-term benefit to myself. In so acting, though, I have not communicated anything about the benefactor to the benefactor, at least if there is no afterlife (or even if there is, but I do not believe in one and hence do not strive to convey any message).

For another case, imagine an indigent offender standing before a judge, awaiting sentence. Suppose that the offender actually hopes that he is sentenced to jail, where he can expect three meals a day, a roof over his head and protection from his enemies, and that the judge does in fact sentence the offender to jail. The offender, thinking the judge is aware of the offender's desire, takes the sentence as an expression of pity or concern on the judge's part and a conveyance of the message that what the offender did was not so bad. However, suppose the judge had no such things in mind. In this case, if the judge was aiming to communicate something to the offender, the judge failed, but this does not undercut the intuition that the judge could be engaged in reactive practices of blaming and punishing, as opposed to acting from an instrumental or objective orientation.

Perhaps the problem with E1 is that it requires communication not only about one who has acted, but also to the one who has acted. Perhaps reactive practices essentially involve communication about a person's action, but not necessarily to the person to whom one is reacting. Hence, consider this principle:

> E2: *A reactive practice is essentially (in part) a communication of one's occurrent attitude about a person's action that is consequent to the action and is directed to someone (not necessarily the one who has acted).*

E2 avoids the counterexamples to E1. In the gratitude case, I was perhaps communicating my appreciation of the benefactor to the benefactor's daughter, and in the punishment case, maybe the judge communicated his disapproval of the offender to members of the courtroom, the police or society at large.

But consider the following counterexample to E2, which suggests that a reactive practice need not involve any communication at all. Suppose that you help me substantially and that I am keen to return the favour. I follow you around, looking for a particularly useful way to repay your kindness, but you do not know that I am doing this. One day an assassin attacks you, and I rush in, taking a bullet that would have killed you and die myself. Because of the surprise to the assassin, you have the time to pull out your gun and shoot the assassin dead before he has time to think about what has happened. You and everyone else alive believe that I was also an assassin who was accidentally killed by my colleague. In this case, I submit that I could be acting gratefully *qua* reactive practice, even though I failed to communicate my sentiments about you to anyone at all.

---

4 For a more careful exposition of punishment as a form of communication in Gricean terms, see Nozick 1981:369-70.

In reply, one might suggest that I have communicated to myself, even if not to anyone else.[5] It does seem as though it is possible to communicate to oneself, e.g., by writing a note to jog the memory of one's future self. However, the present case is unlike that sort of case. It does not seem plausible to construe the action of sacrificing my life for the sake of my benefactor to be a matter of using conventional symbols to transmit a message to myself. After all, I am already aware that I feel grateful to my benefactor and do not need to convey any representations to inform myself of that fact.

The problem with E2 might plausibly be thought that it requires communication of a certain mental state, where 'communication' is a success term, i.e., involves the actual apprehension of a message that someone has sought to convey by using symbols. Perhaps loosening the requirement of success is the next logical step:

> E3: *A reactive practice is essentially (in part) an attempted communication of one's occurrent attitude about a person's action that is consequent to her action and is directed to someone.*

E3 avoids the previous counterexample, since one might suggest that, even if I failed to communicate my thanks to anyone by giving up my life, I was at least trying to do so.

I think, however, that there are several instances of reactive practices that do not include even the attempt to communicate with anyone. Here are two.[6] Suppose, as some philosophers have suggested, that, if God existed, it would be impossible for him to convince us that he does. Perhaps that would be because no conclusive evidence could be marshaled, or because we would not in fact be convinced by it. Regardless, imagine that God knows that it is pointless trying to convey any of his attitudes to us, since we are incapable of understanding his message, and so he does not try. Nonetheless, it seems that God could punish human beings in a reactive way, by smiting evildoers in this world simply because they have done evil and in proportion to the evil they have done.

The case of God is one in which the punishing agent does not try to make his attitude known, while here is a case in which a punishing agent strives to keep his attitude unknown. Taxi drivers in Johannesburg tend to violate the rules of the road, e.g., by running lights, cutting into queues, using turn lanes to go straight and stopping in the middle of traffic to catch a fare. These taxi drivers also tend to be armed. Now, imagine that I follow a taxi driver for a while, taking careful note of how many just laws he has broken and thereby ascertaining how much punishment would be proportionate to his misdeeds. Imagine that I then, in the middle of the night, inflict damage on his taxi to the degree that fits his crimes. Here, in order to avoid retaliation, I aim to keep my identity a secret, and so try to keep my own resentful and disapproving attitude unknown to the driver.[7] Despite the lack of attempt to convey my sentiments, I hardly need to be conceived as acting for consequentialist reasons, and may be understood merely to be imparting what is deserved.

I think that parallel cases could be constructed for the reactive practice of gratitude. Here, too, one could act gratefully to a benefactor for non-consequentialist reasons while striving to remain anonymous. To those who balk at considering this to be *grati-*

5 I doubt I would have thought of this reply, were it not for some remarks by Pedro Tabensky.

6 For another one, see Metz 2000:495.

7 The reader may also imagine that I do not even try to communicate the fact that *someone* is resentful or disapproving of the driver's behaviour; perhaps all I have time to do is give him a flat tire.

tude (or gratitude *qua* reactive practice), suppose that you had two alternatives with respect to a benefactor who has done an enormous amount for you. On the one hand, you could say 'thank you' to her. On the other hand, you could do something really substantial for your benefactor, but doing so would require that she be unaware the good came from you. Now you must choose which course of action to take. Surely, you would take the latter course, of which the best explanation is that you think it would be the better way to behave gratefully from a participant attitude.

A reasonable suspicion at this point is that the problem with the previous three forms of expressivism is that they cash out expression narrowly in terms of communicating or the attempt to communicate. The focus on communication seems responsible for the counterexamples, and, furthermore, it is very hard to see how self-reactive practices such as 'punishing oneself' are to be captured by a communicative version of expression. Fortunately for the expressivist, there are forms of expression besides communication, and, rather than specify them, one option would be to include any of them that might be relevant, as follows:

> E4: *A reactive practice is essentially the expression of one's occurrent attitude about a person's action that is consequent to her action.*

E4 plausibly avoids the counterexamples to E3, for one might suggest that even if God cannot communicate with us, he is expressing his disapproval when he smites evildoers, and that, even if I elect not to try to communicate with the taxi driver, I too am expressing my resentment when I damage his vehicle.

Yet E4 is not invulnerable to strong counterexamples. Consider a judge who disagrees with the state's law and hence believes that the 'offender' before him is morally innocent. Such a judge would lack any ill-will when he sentences this person, but could still be engaging in a reactive practice; the judge need not be thought of as adopting an objective disposition and hence imposing punishment for consequentialist reasons.

In reply, one might suggest that, even if the judge were not expressing a negative emotion or feeling with regard to the one who has offended, he would at least be expressing his belief that the person broke the law or is legally culpable for punishment. This appears hard to deny.

However, an expression of one's belief that someone broke the law or legally warrants punishment is not what Strawson and others[8] have in mind when they deem punishment to be inherently expressive. Strawson does not characterise reactive practices as expressions of mere beliefs, but rather of 'attitudes and intentions' (62, 63) and sometimes of 'feelings' (63), and particularly those of good-will, ill-will and indifference (63, 64, 77). His stock examples of what reactive practices display are resentment, anger, indignation, disapproval, guilt, shame, remorse, esteem and affection, which are hardly well captured as mere judgments. Cognition plays a role in many of these reactions, but even so, in the present context of legal punishment, the belief that someone has broken the law is far from all that the expressivist would say is going on. At the very least, these reactive practices, if essentially expressive of a belief, express the belief that one has *unjustly* broken the law, i.e., exhibit a 'moral attitude' in Strawson's terms (80).

8 Cf Feinberg 1965 and Skillen 1980, who contend that punishment essentially includes the expression of emotions such as contempt, indignation and the like.

Since the reluctant judge does not express any of the non-cognitive or even moralistic attitudes that expressivism is best understood to entail are essential to punishment *qua* reactive, I am led to consider a different version of expressivism. The next point does not conceive of an agent necessarily expressing her own attitudes:

> E5: *A reactive practice is essentially (in part) the expression of someone's occurrent attitude (not necessarily the agent's own) about a person's action that is consequent to her action.*

This version of expressivism avoids the counterexample above, for even if the judge is not expressing his own disapproval, indignation or hostility, he could be using punishment to express these attitudes on behalf of someone else, say, the legislature, state or society.

Even so, I can think of cases in which it appears that a reactive practice does not express any occurrent attitude held by anyone. First, consider a case in which your grandmother gives you socks as a holiday present. Suppose that you really do not appreciate the gift and that you do not even appreciate whatever time and expense your grandmother spent acquiring them. Imagine, though, that you recognise that you ought to feel appreciative in a way you do not, and that you therefore say 'thank you' to your grandmother. In this case, you do not have the attitude normally associated with grateful behaviour, and, by virtue of saying 'thank you', you are also not expressing this attitude on behalf of anyone else. No one is feeling grateful here, and yet you behave gratefully in a way that is not based on the objective attitude.

For a second case, in the context of punishment, consider a variation of the case of the reluctant judge. Imagine that the judge misinterprets a state law, thinking that it requires a certain unjust penalty that it in fact does not. The judge wishes that he did not have to sentence the 'offender' he has found guilty of breaking the law, but thinks to himself that he is paid to give people the punishment they deserve as determined by the state, and moreover has sworn an oath to do so. Here, punishment would not express any of the relevant attitudes on the part of the judge; he is not disapproving, indignant or the like, as he thinks the law is unjust. In addition, although the judge believes that the punishment would express the state's disapproval, indignation, etc., he is incorrect, since he has misinterpreted the law. Hence, punishment in this case would express no negative attitude on the part of anyone, and yet we need not think of the judge imposing punishment for the sake of social utility or any other instrumental function.

What the expressivist should say in reply to these cases is that a practice can express a person's attitude that does not occur at the time that she engages in the practice.[9] Since you normally use 'thank you' to express a feeling of appreciation, perhaps you express this feeling with these words even when you do not actually feel it in the context of your grandmother. And since the judge usually expresses censure with punishment, he might continue to do so in the instance when he punishes without a censorious attitude. So, consider the following version of expressivism:

> E6: *A reactive practice is essentially (in part) the expression of one's attitude about a person's action that is consequent to her action, where one either occurrently has the attitude or is disposed to express it with this practice.*

9 I thank David Martens for pushing me toward dispositional versions of expressivism.

One might wonder what reason there is to favour E6, which says that a reactive practice expresses *an attitude that might be merely dispositional*, over a principle that cashes out a reactive practice in terms of the mere *disposition to express* an attitude. The reason is that saying 'thank you' to your grandmother in fact seems to be an expressive action. It is not as though saying 'thank you' is a case of a disposition to express a feeling of appreciation where the expression is not actual; rather, saying it plausibly does in fact express appreciation, even if the appreciation is not felt at the time.

Counterexamples to E6 will lead me to the most defensible version of expressivism. The trouble with E6 is that it seems possible for a person to express an attitude with a practice even if she neither has this attitude when she engages in the practice nor has used this practice to express the attitude routinely in the past. To see this, modify the case of your grandmother. Imagine that you are a child who does not feel appreciation for the socks, but knows that she is supposed to feel this way. And suppose that you are just learning to say 'thank you' in response to gifts. Here, it seems plausible that you are in fact expressing appreciation when you say 'thank you', even though you did not feel appreciative and have not habitually used 'thank you' in the past to express this feeling. Similar remarks go for a psychopath who never feels appreciative, but mimics the behaviour of people who do.

These cases should seem all the more compelling in light of the old Gricean distinction between speaker meaning, what a speaker intends to say with the words she chooses, and sentence meaning, what a speaker has actually said with them. Just as it is possible for people to say things they do not mean, so it appears possible for there to be 'action' meaning, i.e., for an action to express an attitude that one neither has when performing an action nor has exhibited in the past with this action.

How can a certain reactive practice such as punishment express resentment and disapproval, if the person engaging in the practice does not feel them and has not even routinely used punishment to display such feelings in the past? The answer is presumably parallel to the way a sentence expresses a proposition even when the speaker does not intend to say it with that sentence, viz., very roughly, by speakers in his community having routinely used the sentence (or its parts) to express certain mental states in the past. Although a reactive practice need not express the *performer's* occurrent or even merely dispositional attitude, it could express an attitude that others in his society have been disposed to express with it. So, consider the last and most promising version of expressivism:

> E7: *A reactive practice is essentially (in part) the expression of an attitude about a person's action that is consequent to her action, where someone either occurrently has the attitude or is disposed to express it with this practice.*

E7's reference to 'someone' having the attitude might seem to be vague, but in light of the above counterexamples, it is aptly broad. It appears that a practice can express an occurrent attitude in the performer, a dispositional attitude in the performer, an occurrent attitude in someone other than the performer, or a mere dispositional attitude in someone other than the performer. E7 therefore includes both 'agent' ('speaker') meaning as well as 'action' ('sentence') meaning. As I have suggested, an action can mean something even if the agent does not mean that by the action, by virtue of meaning that has been imbued by societal convention. And, furthermore, an action can express an attitude even if the community has not in the past used the action to express it.

For instance, by virtue of 'agent' meaning, the first time anyone used punishment for retributive reasons, it plausibly expressed resentment or other negative attitudes.

I find it difficult to raise counterexamples to E7. To do so, one would need to find a case in which a reaction is associated with a certain attitude (which is essential to count as a 'reactive practice', as per section 2) but the action neither manifests an occurrent attitude by the agent or someone on behalf of whom he is acting, nor signifies an attitude that the action had often been used to express in the past, whether by the agent or by people in his community. Although I have not yet encountered such a case, I can think of other, more theoretical *prima facie* problems with E7, which I explore in the next section.

4. Questioning the Best Form of Expressivism

In the previous section, I distinguished seven different ways in which one might construe reactive practices as essentially expressive in nature, and argued that one of these conceptions, E7, is preferable to the other six for being able to avoid and best explain counterexamples to them. In this section, I critically discuss some objections that one might raise to E7. As I noted in the introduction, I am unsure whether these objections are sound or not. My goal is merely to articulate and advance discussion about E7, not to settle reflection about whether to adopt it or not.

The first objection to E7 (and all other forms of expressivism) is that it does not cohere with retributive theories of why legal punishment is justified. Our conception of what punishment is *qua* reactive practice ought to cohere with retributive theories of justified punishment, competing accounts of what non-consequentialist punishment by a political community ought to be doing. Retributive theories of punishment are 'backward-looking' views that base the justification of punishment on facts about the past. According to this perspective, whether punishment is justified is a function of whether a crime occurred, and how much punishment is justified is a function of the nature of the crime, such that the worse the crime, the harsher the penalty should be. The correct account of punishment as a reactive and hence non-consequentialist practice ought to be necessary and sufficient to carry out the function of punishment prescribed by a large majority of retributive theories. However, E7 does not fit with them; conceiving of punishment as essentially expressive does not seem to be necessary for punishment to carry out the function prescribed by a large majority of retributive theories, as I now demonstrate.

There are three major forms of retributivism in the literature.[10] According to the desert theory, the point of punishment should be to give a person the harm she deserves for having done wrong. However, giving a person the harm she deserves for having done wrong does not require expressing anything about her action; all it appears to require is harm. In reply, one might suggest that offenders deserve censure; they deserve to be the objects of disapproval and hostility. However, this does not seem true, for in order to give people what they deserve for wrongdoing, it seems sufficient to intentionally impose hard treatment in proportion to, and because of, the person's wrongdoing.

According to the fairness theory, the aim of punishment should be to remove the unfair advantage an offender has taken at the expense of law-abiding citizens. An offender receives the benefits of a state such as protection and order without undergoing his share of the obedience to law required for the state to generate these benefits. In or-

10 I follow the taxonomy employed in Metz 2000, 2004.

der to impose a burden on the offender of the kind and degree that he avoided in breaking the law, it seems that the state need not express anything about him. All it has to do is restrict his will in a way that he did not on his own, but that other, law-abiding citizens did.

Finally, it might seem as though E7 coheres extremely well with a censure theory of punishment, but, even here, this is not obvious. It is true that standard forms of censure theory claim that the reason for the state to punish is to express disapproval of offenders. However, most friends of the censure theory believe that there are deeper reasons to punish, which permit or require this expressive element. In particular, most hold that the ultimate rationales for punishment are to treat offenders as responsible and victims as important, which can or must be done by expressing disapproval of offenders. *Perhaps treating* people in these ways does not require *expressing* anything about them, but instead merely imposing a burden on the offender consequent to, and proportionate to, his action.

One way to reply to this objection would be to 'bite the bullet' and say that the best expressive conception of reactive practices has given us good reason to favour certain forms of censure theory—those that take the expression of disapproval to be *essential* to the justification of punishment—and to reject all other, competing accounts. In a way, I would be happy with this implication, as I have sought to justify the censure theory in a series of papers.[11] Perhaps a good reason to adopt the censure theory over retributive rivals is that the expressive essence of punishment *qua* reactive practice is necessary for realising only censure theory's retributive aim.

However, friends of the desert and fairness theories have a reasonable way to respond here. Perhaps what I have been calling the 'expressive' element of punishment is nothing over and above what I said is necessary and sufficient to realise the aims of desert and fairness. I contended that intentionally imposing hard treatment on someone because of and in proportion to his wrongdoing would give him what he deserves or remove his unfair advantage (or treat an offender as responsible and his victim as important). Maybe so reacting to the offender *constitutes* the expression of disapproval, hostility or whatever negative attitudes are plausibly exhibited by punishment. In other words, perhaps a non-expressive construal of punishment co-refers to an expressive one. Just as a person is referring to H_20 when she speaks merely of 'water', one might be referring to expressive behaviour when one speaks merely of 'intentionally imposing hard treatment on someone because of and in proportion to his wrongdoing'. If this were true, then expressive behaviour would in fact be necessary to carry out a wide array of retributive functions such as realising desert and fairness; it is merely that one need not *call* it 'expressive' behaviour. And hence the expressivist can also reasonably reply to the objection by contending that her account of the reactive practice of punishment does not fail to cohere with the dominant retributive rationales for punishment.

The first objection to E7 is that it and expressivism in general fit poorly with standard retributive accounts of *why* the reactive practice of punishment is morally justified. The next objection is that they fit poorly with common views about the *extent* to which the reactive practice of punishment is morally justified. Most theorists who believe in retributive punishment are universalists, i.e., they believe that certain moral norms, and in particular those governing retributive punishment, apply to people in all societies over time, and not merely to some. However, it appears that conceiving of re-

11 Metz 2000, 2002, 2004, 2006, 2009.

active practices as inherently expressive makes it difficult to avoid relativism about where and when retributive punishment is justified.

Recall the analogy between sentence meaning and action meaning that I drew to bolster E7. I suggested that, just as a sentence could mean something that its speaker did not mean by it, so an action could express an attitude that neither its performer nor anyone else actually has at the time. Now, the way a sentence plausibly obtains a meaning independent of the speaker's occurrent mental states is by a group of speakers having used the sentence (or its parts) to express certain things in the past. In short, sentences obtain meaning by convention, such that, e.g., it would be appropriate to say 'thank you' in order to express gratitude when in the US, but not when in France (where you should instead say '*merci*'). The present objection submits that something similar holds for the expressiveness of actions. If a reactive practice expresses an attitude that neither the agent nor anyone else has at the time, it probably does so in virtue of variable convention. But if variable convention determines what a reactive practice expresses, and if the justification for a reactive practice depends on its essential nature *qua* expressive, then its justification will be a function of variable convention as well. Few theorists believe that the justification of a reactive practice is relative, generally holding that, e.g., retributive punishment is morally justified for all wrongdoers, regardless of the society they live in (at least if there is a political community). Hence, it is unlikely that the essential nature of the reactive practice of punishment is expressive.

To be clear and careful, here is the *reductio* argument against E7 and expressivism generally in standard form:

1. The reactive practice of retributive punishment is justified.
2. If the reactive practice of retributive punishment is justified, it is so in virtue of all its essential properties.
3. The reactive practice of retributive punishment is essentially expressive.
4. Therefore, the reactive practice of retributive punishment is justified in virtue of what it expresses (1, 2, 3).
5. If the reactive practice of retributive punishment is justified in virtue of what it expresses, then it is universally justified only if it is used to express something in all societies (and probably the same thing).
6. Therefore, the reactive practice of retributive punishment is universally justified only if it is used to express something in all societies (and probably the same thing) (4, 5).
7. But it is not true that the reactive practice of retributive punishment is used to express something in all societies (or at least not the same thing).
8. Therefore, it is not true that the reactive practice of retributive punishment is universally justified (6, 7).
9. But the reactive practice of retributive punishment is universally justified.
10. Therefore, it is not true that the reactive practice of retributive punishment is essentially expressive.

Before considering how one might reply to this argument on behalf of the expressivist, let me clarify its elements. In 'Freedom and Resentment', Strawson famously maintains that we cannot avoid believing (1) to be true, and, in any event, as a matter of fact, most non-consequentialists do believe it. (2) is hard to question; for if one sug-

gested that a reactive practice could be justified in virtue of something less than all its essential features, it would seem more plausible to contend that what would be justified would instead be whatever would remain if one subtracted those features from the reactive practice that do no justificatory work. (3) is simply the doctrine of expressivism. (4) follows logically from the first three premises. (5) seems plausible in light of reflection on the analogous conditions under which it is appropriate to say a certain sentence. If it is appropriate to say a certain sentence in virtue of what it expresses, then it would be apt to say a certain sentence in all societies only if it had a meaning in all societies, and probably only if it had the same meaning in those societies (it would be quite a fluke if a certain sentence meant a variety of different things from culture to culture). Similar things go for an action that has meaning; it is apt to perform it in virtue of what it expresses in all societies only if it expresses an attitude, and probably the same attitude, in all societies. (6) follows logically from the previous two premises. (7) is an empirical claim to the effect that not every society has employed retributive punishment[12] and that, even if every society has done so, societies have assigned different meanings to the practice. Again consider an analogy with sentences. As far as I know, there is only one word that has the same meaning in all cultures that use the word, namely, 'Amen'. Just as one quite usually encounters different words in different cultures, and different meanings ascribed to the same words in different cultures, so it is plausible to think that one will encounter different actions in different societies, and different meanings ascribed to the same actions in different societies. (8) follows logically from the previous two premises. (9) is an uncontroversial claim among the friends of reactive practices. If you believe that reactive practices are justified, then you probably believe that they are universally justified, e.g., that, in the case of retributive punishment, it would be wrong for all political societies to punish the innocent or acquit the guilty for the sake of marginally good consequences. As the first nine premises have generated a contradiction, one of them must be false. The objection is that the culprit is (3), the doctrine of expressivism, which seems more controversial than the other premises.

One way to reply to the objection is to point out that (2) is not as uncontroversial as it seems. It appears that a policy can be justified (and on non-consequentialist grounds) for something less than all its essential features. For instance, suppose that affirmative action were justified on grounds of compensatory justice. The essence of affirmative action is a matter of awarding an opportunity to someone in part because she is not a white male (where not being a white male is not a qualification for the opportunity). Now, such a practice could be justified because of a need to compensate non-white males for past injustice, even though it is the case that sometimes affirmative action will not serve that function. Some white males will not have been responsible for or benefited from past injustice, and some non-white males will not have been harmed by it. Nonetheless, a policy of affirmative action might be the best that a state is in a position to do in order to realise compensatory justice. A state will not have the resources that would be required to ascertain precisely who is entitled to how much compensation; an essentially race-based approach might be an approximately accurate way of effecting redress and be justified by virtue of the lack of any more accurate and

12 In particular, traditional African cultures are known for eschewing retributive conceptions of the function of punishment, instead generally favouring forward-looking rationales such as the need to protect the community from the wrath of angry ancestors or to foster reconciliation between the offender, the victim and the rest of the community.

realistic alternatives. Furthermore, suppose that the adoption of affirmative action would express concern on behalf of the political community for victims of past injustice. Even if affirmative action essentially expressed this, one might argue that it is only the compensatory function, and not the expressive one, that justifies it. Hence, affirmative action could be justifiably adopted for a certain (backward-looking) rationale, even though (in two respects) not all of its essence performs the function prescribed by this rationale.

Something similar could hold for punishment as a reactive practice. Even if part of the essence of retributive punishment were to express certain attitudes, it might be that such punishment would be justified for (non-consequentialist) reasons other than those relating to its expressiveness. The expressiveness might merely 'come along for the ride' with whatever does the justificatory work. Note, however, that this reply appears to be in tension with the reply to the first objection, where it seemed best for the expressivist to say that expressive behaviour is nothing distinct from the 'non-expressive' behaviour of intentionally imposing hard treatment on someone because of, and in proportion to, his wrongful deed.

Another way to reply to the second objection is to question (7). So far as I am aware, it is a fact that not all societies have used retributive punishment. However, one may question whether a practice must actually be employed in order for a society to understand it to express a certain attitude. On the individual level, I might never forgive others, but I could well understand that others who forgive express a kind of good-will (that I do not have or choose not to exhibit). Analogously, even if a society does not use punishment as a reactive practice, it is hard to see how it could think of retributive penalties as expressing something other than disapproval, hostility and the like; surely, no society would think that punishment *qua* reactive is an expression of indifference, let alone affection or some other form of good-will. Just as it is probably universally accepted that tears mean sadness and keeping one's distance expresses dissatisfaction, so all societies might believe that retributive punishment signifies moral disapproval proportionate to injustice, even if some of them lack this attitude or choose not to express it.

5. Conclusion

I am frankly unsure what to conclude with regard to whether expressivism is true or not. I have distinguished seven different versions of the view that reactive practices are essentially expressive of moral attitudes, and argued that one of them entails and explains our intuitions about their nature better than the others. The favoured conception, E7, denies that expressive behaviour must be communicative or must express the agent's own occurrent or dispositional attitudes—or even the occurrent attitudes of anyone at all. Instead, if a reactive practice is inherently expressive, it can be so in non-communicative ways that display attitudes that no one actually has at the time of its instantiation. I noted that one objection to this view, and to expressivism as such, is that any expressive function inherent to the reactive practice of punishment seems irrelevant to its justification, in light of most major retributive theories. The expressivist could reply that expressive terms co-refer with non-expressive ones, such that whatever property the objector thinks does justify punishment is nothing separate from the property that the expressivist believes is expressive. I find this reply *prima facie* worth taking seriously, but I am not convinced that it is correct. I noted that a second objection to E7 and expressivism in general is that, if the reactive practice of punishment

were essentially expressive, it would be justified in virtue of what it expresses, but that what a practice expresses would be a function of variable convention, meaning that expressivism cannot capture the universal scope of justification that most retributivists believe is true of the reactive practice of punishment. One reply on behalf of the expressivist is that a practice might be justified by virtue of only some of its essential properties, and not all of them (in particular the expressive ones), but I am not sure that the analogy of affirmative action that I used in support of this is strong. Another reply on behalf of the expressivist is to claim that, even if what the reactive practice of punishment expresses depends on convention, all societies as a matter of fact believe that it expresses the same thing, even if they have not used the practice themselves. Again, I am currently unable to judge how plausible this claim is. Although I cannot draw a firm conclusion about whether expressivism is true or not, I have specified which version is most likely to be true, and have indicated some *prima facie* problems facing it that must be addressed in future work in order to make a belief in expressivism reasonable.

References

Feinberg, J. 1965. 'The Expressive Function of Punishment' *The Monist* 49: 397-423.

Metz, T. 2000. 'Censure Theory and Intuitions about Punishment' *Law and Philosophy* 19: 491-512.

Metz, T. 2002. 'Realism and the Censure Theory of Punishment' *Archives for Philosophy of Law and Social Philosophy* 85: 117-29.

Metz, T. 2004. 'Legal Punishment' in C. Roederer and D. Moellendorf, eds. *Jurisprudence*. Lansdowne: Juta and Company, Ltd, pp. 555-87.

Metz, T. 2006. 'Judging Because Understanding: A Defence of Retributive Censure' in P. Tabensky, ed. *Judging and Understanding: Essays on Free Will, Narrative, Meaning and the Ethical Limits of Condemnation.* Aldershot: Ashgate Publishing Ltd., pp. 221-40.

Metz, T. 2009. 'Censure Theory *Still* Best Accounts for Punishment of the Guilty: Reply to Montague' *Philosophia* (forthcoming).

Nozick, R. 1981. *Philosophical Explanations*. Cambridge: Harvard University Press.

Skillen, A. 1980. 'How to Say Things with Walls' *Philosophy* 55: 509–23.

Strawson, P. F. 1962. 'Freedom and Resentment' repr. in G. Watson, ed. *Free Will.* New York: Oxford University Press, 1982, pp. 59-80.