8

Responsibility Beyond Blame: Unfree Agency and the Moral Psychology of Criminal Law's Persons

CRAIG REEVES*

> We need to recast our ethical conceptions … not in order to escape or adjust ourselves to determinism or naturalistic explanation … [but] in order to be truthful even to what we know already about our psychology.[1]
>
> If it is true of any institution that it has inscribed into it a psychology, this is likeliest to be … the criminal law.[2]

I. INTRODUCTION

IT IS OFTEN claimed that a therapeutic stance towards offenders and respecting them as persons are incompatible: only the retributive or 'just deserts' model, and not the 'treatment' or 'therapy' model, can respect offenders as persons. Treatment or therapy inevitably disrespects offenders, seeing them as less than persons, so criminal justice should be self-consciously oriented to holding responsible, attributing blame and administering punishment according to an abstract and thus impartial set of standards relating to the culpability of conduct. It should self-consciously refrain from a therapeutic stance that takes criminal offending to be a manifestation of some kind of pathology. Therapeutic or rehabilitative interventions are permissible but only as a supplement to retribution, for a departure from retributive reason is bound to entail

* I am very grateful to Alec Hinshelwood, Jaakko Nevasto, Alan Norrie and Matt Sinnicks for their helpful comments on earlier drafts as well as for many excellent conversations on the issues. I am also indebted to Matt Matravers for his insightful comments on and encouragement with the chapter.

[1] B Williams, *Making Sense of Humanity* (Cambridge University Press, 1995) 19.

[2] R Wollheim, *The Mind and Its Depths* (Harvard University Press, 1993) 129.

a denial of the offender's agency and a suppression of their capacity for rights, that is, a denial of the respect owed to them as a person.

In a series of recent papers, Nicola Lacey and Hanna Pickard have challenged this orthodoxy, arguing that, on the contrary, therapy can be reconciled with respect for persons, because we can reconcile therapy with the justice model. They argue that Pickard's 'clinical model' of 'responsibility without blame'[3] based on reflection on therapeutic work with psychopathology shows us how to reconcile the justice model of criminal responsibility and punishment with the reparative and reconciliatory ideal of a therapeutic dialogue animated by a spirit not of blame but of forgiveness.[4]

The retributivist argument that only retributive punishment involving hard treatment in proportion to wrongdoing can respect offenders as persons got considerable mileage out of a stereotype of therapy: that it involves reifying the patient, ceasing to engage with them as a person, and treating them as a mere thing, outside the 'moral community'.[5] Lacey and Pickard insist this stereotype is ignorant of the facts: therapists working with psychopathology of the sorts common among offenders in actual fact do hold patients responsible for their actions and impose consequences, thus deploying the basic norms of the justice model, and they do this as an integral part of their therapeutic practice. But they do this not in a retributive spirit, but from a compassionate stance of therapeutic concern that involves avoiding the blaming emotions.

The clinical case shows that holding responsible for actions and imposing consequences for wrongdoing – 'detached blame' – and the negative moral emotions and the imposition of suffering or hard treatment – 'affective blame' – are separable. The norms of detached blame say essentially that 'responsibility tracks agency', so that people are responsible where their agency was appropriately engaged,[6] but affective blame is distinct from and not entailed by such judgements of detached blame. This distinction is actual in therapy, and if it is actual it is possible, so retributivism is wrong: a criminal process that both holds responsible according to the justice model and thus respects offenders as persons, and approaches punishment in a forgiving and reconciliatory way, is possible.

[3] H Pickard, 'Responsibility Without Blame: Empathy and the Effective Treatment of Personality Disorder' (2011) 18 *Philosophy, Psychiatry and Psychology* 3; H Pickard, 'Responsibility Without Blame: Philosophical Reflections on Clinical Practice' in KWM Fulford et al (eds), *Oxford Handbook of Philosophy and Psychiatry* (Oxford University Press, 2013).

[4] N Lacey and H Pickard, 'From the Consulting Room to the Courtroom? Taking the Clinical Model of Responsibility Without Blame into the Legal Realm' (2013) 33 *Oxford Journal of Legal Studies* 1; N Lacey and H Pickard, 'To Blame or to Forgive? Reconciling Punishment and Forgiveness in Criminal Justice' (2015) 35 *Oxford Journal of Legal Studies* 665.

[5] P Strawson, 'Freedom and Resentment' (1962) 48 *Proceedings of the British Academy* 1, reprinted in J Fischer and D Ravizza (eds), *Perspectives on Moral Responsibility* (Cornell University Press, 1993) 59.

[6] Lacey and Pickard, 'To Blame or to Forgive?' (n 4) 666.

Further, we not only can but ought to institutionalise a model of criminal justice that places forgiving and reconciliatory dialogue, rather than blame and hard treatment, at its core. Although we hold responsible and punish in accordance with the justice model, we need not and should not do so in a spirit of affectively blaming. We can and should aim for criminal practices informed by the reparative forgiveness strategies of communicating the harm done by the offender's behaviour, indicating the potential for repair and maintenance of a valuable relationship (between offender and community) if wrongdoing is avoided in future, and reminding the offender of the importance to them of the valuable things that are at stake in maintaining that relationship.[7] Guided by these, we should eschew punitiveness and put reparative and reconciliatory strategies at the centre of criminal punishment.

This is a richly suggestive account animated by a humane impulse, and it challenges the retributivist revival to take seriously the real psychology of criminal law's persons. Yet to me it seems that Lacey and Pickard's position does not go as far as – according to certain of its own impulses and insights – it ought to in challenging retributivist dogma and our existing criminal practices.

It is unclear, after all, how different their proposals are from those of prominent humane retributivists like Antony Duff. They insist that 'hard treatment' is not essential to criminal punishment, but 'serious consequences', 'no doubt typically negative but occasionally not',[8] are, and it is not clear how deep this distinction goes. It is the stigma, exclusion and condemnation of retributive hard treatment that they oppose, but humane retributivists like Duff are opposed to those as well, construing punishment as about repentance, reparation and reconciliation.[9] They seem to assume that condemnation, stigma and exclusion are primarily a function of 'affective blame', so that if we jettison affective blame while keeping detached blame, this would allow us to remove these painful features of punishment while retaining the basic structure of our practice of holding criminally responsible and punishing. But this is controversial. If it seems unlikely that hard treatment will be able to avoid stigma, exclusion and condemnation, it seems equally unlikely that 'serious negative consequences' imposed in response to criminal conviction within our existing practice will, for in virtue of being so imposed, and thereby of meaning what they mean, they are likely to be painful in the same ways.[10] Stigma and exclusion are to some extent bound to attach to any punishments within our existing criminal justice practices, to any 'serious consequences' imposed for criminal conviction.

[7] ibid 680–81.

[8] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 8.

[9] RA Duff, *Punishment, Communication, and Community* (Oxford University Press, 2001) 106–09.

[10] See J Feinberg, 'The Expressive Function of Punishment' (1965) 49 *The Monist* 397, 397–423.

Perhaps Lacey and Pickard overestimate the importance of 'affective blame' – the blaming emotions – in retributivist theory and practice. While some – notably, expressivist – versions of retributivism do explicitly see punishment as properly expressing blaming emotions like anger, resentment, indignation or hatred,[11] most do not make affective blame central to punishment at all. For Duff, for example, punishment's point is not to express negative emotions but to communicate moral criticism in a rational dialogue, and through that dialogue to encourage reparation and reconciliation. Even for retributivists who, like Michael Moore, give the blaming emotions a central place in the justification of retributivism,[12] punishment is not itself about expressing those blaming emotions; its point is to give effect to the moral demands that our moral emotions reveal. And in actual practice as well it is unclear that affective blame plays much of a role in retributive punishment.

This might explain why, when we jettison affective blame and keep detached blame, we are still left with something in the ilk of humane retributivism. Lacey and Pickard's model remains caught in the unstable tension that characterises our existing criminal practices: the real psychology of criminal law's persons, the limitations to freedom that can in fact arise given our psychology, and the social constraints that shape and limit our capacities, are taken into account at the sentencing and punishment stage, but these same features are suppressed at the conviction stage of criminal law's judgement of responsibility, 'detached blame'. Lacey and Pickard accept this compromise even while certain of their insights speak against it.

This is, I suspect, because they concede a crucial retributivist premise: that respect depends on holding responsible,[13] 'detached blame'; that anything else might reify or 'dehumanise' offenders,[14] treating them as mere things or animals rather than rational agents. That premise, though, presupposes a theory of agency and a moral psychology, both questionable, which are embedded in our existing holding responsible practices. Lacey and Pickard's acceptance of this retributivist premise prevents them from getting into view something that their own reflections on affective blame imply: that our moral emotions disclose the inappropriateness not only of punitiveness but of the categorial structure of our practice of holding criminally responsible, 'detached blame', itself.

Far from being uniquely placed to respect offenders as persons, the retributive practice of detached blame, of holding responsible and punishing, is

[11] See J Feinberg, 'The Expressivist Function of Punishment' in J Murphy and J Hampton (eds), *Forgiveness and Mercy* (Cambridge University Press, 1988). On expressivism, see C Reeves, 'What Punishment Expresses' (2019) 28 *Social and Legal Studies* 31, 31–57.

[12] MS Moore, *Placing Blame: A Theory of the Criminal Law* (Oxford University Press, 1997) chs 1–4.

[13] V Tadros, 'Poverty and Responsibility' (2009) 43 *Journal of Value Inquiry* 391.

[14] J Gardner, 'On the General Part of the Criminal Law' in RA Duff (ed), *Philosophy and the Criminal Law* (Cambridge University Press, 1998) 254.

incapable of doing so. Therapy can be reconciled with respect, not because therapy can be reconciled with the justice model, but because respect does not depend on the justice model at all. The justice model's austere conception of respect ignores the psychology of real persons and fictionalises them as autonomous beings abstracted from the constraints, limitations and privations of psychological reality. Lacey and Pickard's model is unable to follow through on its important insights into the normative significance of the real psychology of persons because it remains committed to a responsibility practice governed by such a fiction. In order to institute respect for criminal law's persons, a radically transformed responsibility practice would be necessary. Or so, at least, I shall argue. I begin by arguing that what undermines affective blame must undermine detached blame as well (section II), and then develop an account of what that might be (sections III–V), before considering the implications of this for our criminal responsibility practice, and some of the reasons why these might be resisted (sections VI–VII), concluding with a suggestion as to how to proceed from here (section VIII).

II. BLAME AND BLAMEWORTHINESS

The claim that we can and should preserve detached blame while rejecting affective blame, thereby reconciling the 'justice' – that is, retributive – model of detached blame with forgiving punishment presupposes that 'detached' and 'affective' blame are normatively separable: that there are phenomena which ground reasons which rationally undermine affective blame that do not also count against detached blame (where by 'rationally undermining' I mean reasons which are both justifying and psychologically efficacious). Only if this is so can it be consistent to disavow affective blame while endorsing detached blame, for only then can there be grounds which can move us to do so whose moving us is owed to their justifying us in doing so. And my immanent criticism of the model is quite straightforward: that this premise of normative separability turns out on inspection to be false. So in fact that which rationally undermines affective blame also rationally undermines detached blame.

Lacey and Pickard offer three kinds of reason said to rationally undermine affective blame, which correspond to the three rationality-conditions on blaming emotions that Pickard has elsewhere proposed,[15] and none satisfy the normative separability test. First, they propose subjective reasons, that is, reasons which focus on us and our wider aims and purposes. Basically, affectively blaming and the punitive practices it motivates in criminal justice will not do any good and may even frustrate our wider purposes – reduction of offending and public

[15] H Pickard, 'Irrational Blame' (2013) 73 *Analysis* 613, 624.

protection, rehabilitation and reintegration of offenders into the community, say.[16] Drawing on work in evolutionary psychology, they point out that

> [r]etributive punishment that stigmatises and gives license to expressions of affective blame may therefore serve to further alienate such offenders from society – in effect, increasing the divide between 'us' and 'them' and shifting an already marginalised and underprivileged faction of our community into a bona fide out-group, thereby confirming their belief that there can be no valuable relationship between society and them.[17]

If we want to reduce reoffending and protect the public, the best strategy is a reintegrative and reconciliatory approach to punishment through forgiveness rather than affective blame. At the same time, detached blame – holding criminally responsible – is claimed to actively further these wider aims.

These subjective reasons correspond to Pickard's condition on rational blame that it 'must not actively undermine rational ends'.[18] They are essentially strategic reasons concerning what it is instrumentally rational for us to feel and do given our wider aims. This sort of criticism of affective blame fits within the broader tradition of what Srinivasan has recently called a 'counterproductivity critique'.[19] But as she points out, subjective, strategic reasons are merely extrinsic reasons for not feeling or expressing some emotion, which are tangential to the question of whether those emotions are intrinsically appropriate.

Indeed, it is odd that Pickard counts this sort of subjective strategic rationality as among the rationality-conditions for the moral emotions at all, for we can ask of anything whether it conflicts with our wider rational ends, but we would not entertain 'conduciveness to our wider rational ends' as one of the rationality-conditions of belief, say: we cannot, and generally should not even if we could, believe according to what suits our wider aims. Revising a belief just because holding it conflicts with our wider aims is a central case of doxastic irrationality: rationalisation.

Strategic rationality seems similarly to miss the point with the blaming emotions as well – as Srinivasan puts it, 'a shift of focus from intrinsic to instrumental justification for anger often comes across as a non sequitur (at best) and morally obtuse (at worst)'.[20] Though one can always ask about how anything fits with our wider aims, it is unclear that such extrinsic considerations properly belong to accounts of the rationality of rational phenomena. They appear to support the separability of the reasons bearing on detached and on affective blame only by side-stepping this more fundamental issue. That is, subjective,

[16] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 21.
[17] ibid 22.
[18] Pickard, 'Irrational Blame' (n 15) 624.
[19] A Srinivasan, 'The Aptness of Anger' (2018) 26 *Journal of Political Philosophy* 123, 125.
[20] ibid 128.

strategic reasons do not seem to be the right kind of reasons for not affectively blaming.[21] In light of this it is unsurprising that, as Lacey and Pickard admit, such strategic considerations are not likely to cut the motivational mustard with the blaming emotions. We may see strategic reasons not to blame and yet remain, and rightly so, unmoved, because they are not reasons of the relevant sort.

Second, they propose intersubjective or relational reasons not to affectively blame, corresponding to Pickard's condition that 'blame must be appropriate to the nature of the relationship'.[22] Given the hardships many offenders have faced in life (trauma, abuse, neglect and deprivation of various kinds) that are typically connected to social injustices or failures of the community, we may think that

> [w]hen children grow up in our midst subject to such conditions, arguably we as a society bear some responsibility for the harm inflicted on them if we fail to intervene. Our responsibility, in turn, may undercut our moral standing to affectively blame the adults these children become.[23]

As a political community, our complicit responsibility deprives us of the standing to express or even feel affective blame toward offenders; the normative character of our relationship as a community to those of our members we have failed gives reasons in political morality against affective blame. But these, too, fail to underwrite normative separability: for it is unclear why our community's responsibility and complicity don't similarly undermine our standing to hold people responsible and punish them altogether. Is it not the case that, as Murphy argued, since 'just punishment rests upon reciprocity', and a feature of 'most existing communities [is] the absence of such reciprocity', that 'punishment is unjust in such a setting because it involves pretending (contrary to fact) that the conditions of justified punishment are met'?[24] And if we ask, in Duff's words, whether the community has 'the right to call [offenders] to account for their wrongs, with suitably clean collective hands and with clear consciences',[25] why should the same factors that allegedly vitiate affective blame not compel us to answer 'no'?

They assert that 'the appeal to adverse early environment and social inequality does not eliminate criminal responsibility or argue against accountability' because 'responsibility is attributed simply in virtue of agency',[26] but this is changing the subject. The relevant question here is not what the grounds of

[21] See P Hieronymi, 'The Wrong Kind of Reason' (2005) 102 *Journal of Philosophy* 437, 437–57; O Na'aman, 'The Fitting Resolution of Anger' (2020) 177 *Philosophical Studies* 2417.

[22] Pickard, 'Irrational Blame' (n 15) 624.

[23] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 24.

[24] JG Murphy, *Retribution, Justice and Therapy* (Reidel, 1979) 80.

[25] RA Duff, *Answering for Crime: Responsibility and Liability in the Criminal Law* (Hart Publishing, 2007) 192.

[26] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 24.

attribution of responsibility are, but whether we as a community have the standing to hold persons to account, to hold them responsible, and punish them, at all, given what Lacey and Pickard see as our community's complicity with the offender's hardships and offending. Indeed, they seem to concede the point when writing that 'the moral standing to hold to account is also arguably premised on relatively equal relationships, and is hence undermined in radically unequal societies such as ours'.[27] The obvious reply might be to insist, as Duff did in an earlier piece, that although our community is unjust and fails people, it is not so bad that it undermines our standing to hold them responsible.[28] This would save detached blame, but it will not underwrite normative separability, for the obvious rejoinder would be that if things are not so bad, they are presumably not so bad that we cannot affectively blame, either. For it is unclear why our standing to hold responsible at all should be any more resilient in the face of radical community failures than our standing to affectively blame. On normative separability, the intersubjective reasons fare no better than the subjective ones.

Now Lacey and Pickard accept that 'just as clinicians no doubt sometimes fail to keep affective blame at bay' despite the good subjective, strategic reasons to do so, in criminal justice as well the subjective reasons not to affectively blame will not necessarily be sufficient to motivate not affectively blaming, and suggest another strategy they think helps in clinical work and can do so in criminal justice as well: reflection on 'the whole of the person and the whole of their story',[29] the person's whole reality. This takes in reflection on that person's present psychological and social situation, and on their past experiences, life history. Reflection on a person's life narrative will typically reveal, in clinical and criminal justice contexts, histories of 'severe childhood psychosocial adversity'[30] and other social hardships and exclusions, connected to class-based, racialised and gendered domination and exclusion. Hence we should see the offender 'not only as one who harms, but as one who has been harmed', as 'both perpetrators and as victims', and doing so will help to motivate not affectively blaming: 'at least reducing, if not outright extinguishing, its force'.[31]

Now to the extent that reflection on the person's whole reality rationally undermines blaming emotions, that can only plausibly be because such reflection reveals objective reasons not to affectively blame – that is, reasons bearing on the intrinsic appropriateness of blaming emotions toward the person who is their potential object, in the whole context. Such reasons correspond to Pickard's

[27] ibid.

[28] RA Duff, 'Law, Language and Community: Some Preconditions of Criminal Liability' (1998) 18 *Oxford Journal of Legal Studies* 189. By 2007 Duff's position is more equivocal, see Duff, *Answering for Crime* (n 25) 192–93.

[29] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 23.

[30] ibid 24.

[31] ibid 23.

third (objectivist) requirement, that the person blamed 'must be blameworthy and so justly deserve a hostile, negative response',[32] and it is now clear what the problem is: the objective rationality-condition for affective blame just is blameworthiness, that is, desert of detached blame, of being held responsible. The crucial consideration is the irreducible question of the intrinsic fittingness of the blaming emotions to the person who is their object, but this is the very same question as the question of detached blame, that is, of the intrinsic appropriateness of holding responsible itself. The responsiveness of our blaming emotions to reflection on the person's whole reality is an implicit judgement about blameworthiness itself, and thus puts in question not only blaming emotions but detached blame – holding criminally responsible – itself.

III. THE TYRANNY OF THE PAST

We might think the ethically relevant feature of the whole reality of the person is not the deprivations and hardships in the person's past per se, but rather the effects that those experiences have had on the person they now are. Lacey and Pickard, though, reason backwards from those effects rather than putting them centre stage: 'given the degree of psychiatric morbidity in the prison population, it is reasonable to conclude that many … are not only perpetrators, but also past victims'.[33] When we reflect on the often terrible hardships, trauma and suffering that is a typical feature of both patient and offender histories, this provokes 'compassion and understanding', and thereby dislodges blame.

Now, that someone has suffered in the past might make us feel compassion, but that does not on its own rationally undermine blame. The suggestion here might be we feel they have 'suffered enough'. But as retributivists have often pointed out,[34] this inference seems implausible. In itself, past suffering does not cancel present blameworthiness. We do not blame someone less on finding out they had flu recently. Alternatively, it might be that such reflections motivate not affectively blaming because of some general determinist incompatibilist commitment: our blaming emotions subside in response to the thought that their crimes were caused (by their past suffering). But as compatibilists have argued, that thought would not account for the specific focus on offenders' traumatic pasts. If everything is causally determined, and that fact undermines affective blame, then that should hold across the board, and the fact of psychosocial hardship should be neither here nor there.[35] Neither interpretation, then, can rationally explain the impact of someone's past hardship on our blaming emotions.

[32] Pickard, 'Irrational Blame' (n 15) 624.
[33] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 24.
[34] eg, G Sher, *Desert* (Princeton University Press, 1987).
[35] eg, Moore (n 12) 487.

How, then, does the person's traumatic past bear on our moral emotions? It must be mediated by our appreciating not just that someone has suffered, but that their past suffering distinctively explains how they came to be as they now are. It is the link between their past and their present reality that is crucial. While Lacey and Pickard reason back from 'psychiatric morbidity' to presumed past victimhood, this gets the cart before the horse: what is important is not as such that they have been victims of harm, but the consequences of such past harms for their present psychological actuality. Lacey and Pickard gesture towards this thought when emphasising that such offenders 'have not only suffered terrible harm, but … have also not been given the opportunity to learn how to behave as moral citizens should'.[36] This situates the importance of the person's history in relation to a deeper understanding of who they have become and why they have acted as they have. Against this, Gary Watson suggests that when we only know about a wrongdoer's present situation, we have no grounds for not blaming them, but when we learn about their past and find a history of abuse and deprivation, we change our understanding: we now make sense of why they are that way, for anyone – we might think – who'd had their life might well have become much like them.[37] But this is to separate the person's present reality from their past in an unsustainable way, flattening out the different possible modes of our relation to our past and of its bearing on who we are in the present. In thinking that the person's present situation gives us no grounds for not blaming, Watson assumes that the present person is a product of their past only in the sense that anyone is, in a truistic and empty sense, a product of their past. In seeing no difference between this general sense in which we are all products of our past, and the distinctive ways in which someone who has suffered a life of neglect, trauma or abuse may be a product of their past, Watson is pushed towards thinking of the significance of a traumatic past as having to do with the general significance of an event-determinist metaphysics for responsibility.

Lacey and Pickard's remarks about the significance of the past lead them in a similar direction: the effect of a traumatised offender's past on them is not in kind different from the effect of anyone else's; it is just that their history has led them to being the kind of person who does not know how to be well behaved. This view is inevitable if one assumes a certain implausible but widely accepted event-based metaphysics, but it is unviable because it collapses the distinction between the influence of the past and the tyranny of the past.[38]

As MacIntyre has recently pointed out, a view on which every 'adult life is equally shaped and controlled by the past' is one in which 'some of the crucial differences' between agents with certain sorts of psychopathology and rationally

[36] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 24.

[37] G Watson, 'Responsibility and the Limits of Evil' in G Watson (ed), *Agency and Answerability: Selected Essays* (Oxford University Press, 2004).

[38] See R Wollheim, *The Thread of Life* (Cambridge University Press, 1986) ch 5.

healthy agents are 'obliterated'.[39] The sorts of traumatic life-histories typical of serious offenders can cause people deep and lasting psychological harm. It is not know-that but know-how that they've been denied the chance to learn: know-how consists not of propositional content, but of rational powers (or capacities). The developmental disruption, disturbance or privation of certain central rational powers is what is picked out by the concept of the tyranny of the past as distinct from the mere having of a past. Someone's traumatic past rationally bears on our emotional response because it deepens our understanding of their present by making comprehensible how they might be in the present tyrannised by what they have in the past undergone.

The relevance, then, of reflection on someone's past history lies in contextualising and deepening our understanding of the person they have become and the ways they have come to act. It is about not merely seeing the offender 'not only as one who harms, but as one who has been harmed', but seeing them as one who harms because they have been harmed, where the explanatory 'because' flows not through indiscriminately causally deterministic chains of events, but through the developmental privation – inhibition, suppression, distortion – of the person's rational powers.Life history is relevant to the extent that past experiences have blocked, distorted or undone the acquisition or development of rational powers that are essential to agency and necessary for what we may call, in the language of moral psychology, rational health,[40] or in the language of moral philosophy, freedom, in the sense of autonomy. This is why attention to a person's past is important. My suggestion is that it is by generating an inchoate appreciation of this unfreedom or heteronomy manifest in many cases of psychopathology and in many others that would not be clinically so described, that reflection on the person's whole reality rationally undermines affective blame by undermining its objective aptness, that is, by undermining blameworthiness.

IV. UNFREE AGENCY

In denying this, Lacey and Pickard must claim that the agentive capacities the norms of responsibility supposedly track are properly engaged even while affective blame would be inapt. They are the familiar 'volitional capacities' of 'choice and a sufficient degree of control', and those 'cognitive capacities' such that '[the person] know[s] what they are doing when they commit an offence'.[41]

The immediate problem here is that the concepts 'choice', 'control' and 'knowledge' are equivocal as to their referent, and this obscures the fact that

[39] A MacIntyre, *The Unconscious*, rev edn (Routledge, 2004) 9.

[40] See R Moran, *Authority and Estrangement* (Princeton University Press, 2001); E Harcourt, 'Containment and Rational Health' (2017) 26 *European Journal of Philosophy* 798.

[41] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 2.

what is implicitly required by our holding responsible norms is something more than agency simpliciter: it is autonomous agency.

It is in one sense trivially true that agents 'exercise choice and control' and 'know what they are doing' in any particular action: if it is an action, something they do rather than something that merely happens to or in them, there are at least various choices the agent makes (whether reflectively or not) concerning precisely how to move their body, they exercise control over their body as it moves, and they know what they are doing under some true description. This is what is involved in the freedom of spontaneity, analytic to agency.[42] But this does not entail autonomy – choice, control and knowledge in the ethically relevant sense.

For choice over aspects of the particular action does not entail choice over the kind of action. And control of one's body does not entail control of one's actions: one may 'control one's body', that is, move one's body, without being in control of the kinds of actions one does. That requires agency with respect not merely to one's movements but to one's ends. But it is surely the kind of action, not the particular action, that matters for blameworthiness.[43] Spontaneity powers are not unimportant because they are essential to agency, but the engagement of autonomy powers, rational agency with respect to one's ends, is what counts for blameworthiness. And knowledge of what one is doing under some description is necessary to agency, but that minimal sort of knowledge does not entail the right kind of practical knowledge at the right level of description: the level of the end that is practically efficacious in it. That may not be known to the agent, either in the right way, or at all.

The sort of knowledge relevant here is a special kind of self-knowledge that is essentially first-personal, and the agent may not know the end that is practically efficacious and thereby realised in their action in this first-personal way. They may find their action, or the end they are pursuing in it, unintelligible; or they may find their act intelligible but only because they have distorted their apprehension of the world in which they act, so that the morally salient true descriptions under which they are acting are unavailable to them. In either case, they may properly be said to not know what they are doing in the relevant way, despite 'knowing what they are doing' in the minimal, spontaneity sense.

The kinds of choice, control and knowledge that is analytic to agentive spontaneity, then, do not entail autonomy. Persons are not necessarily autonomous; their actions may be fully actions, and yet unfree, heteronomous. And this – I want to suggest – is what is presupposed by, and explains, the way our moral emotions are affected by reflection on the person's whole reality.

Autonomy of action means that the action is the agent's own, but over what this amounts to there is deep dispute. Subjectivist accounts of autonomy

[42] H Steward, 'Fairness, Agency and the Flicker of Freedom' (2009) 43 *Noûs* 64.
[43] As Steward argues in 'Fairness, Agency and the Flicker of Freedom' (n 42).

understand autonomy to be something to do with the internal coherence-relations within a person's psychological items or 'motivational set', and which connect autonomy to the phenomenological, such that it is constitutive of heteronomy that it shows up in the person's experience, as felt alienation. The most influential such 'hierarchical' accounts stipulate that autonomous actions are those flowing from effective desires (which move the agent to action) the agent is identified with, where that means the desires cohere with some special 'higher-order' mental items that 'speak for' the agent ('higher-order desires',[44] 'values',[45] the desire to be rational,[46] temporally projected 'plans and policies',[47] etc).

Such accounts turn out be inadequate to both our pre-theoretical intuitions and clinical practice. They insist that autonomy is purely internal to the psychology of a person so that only failures that show up phenomenologically for the agent can count as failures of autonomy, but that seems to fail to account for various cases where we would want to deny the person is autonomous, such as systematic brainwashing, sci-fi mental-state-manipulation,[48] or psychotic delusion. In such cases, that the person does not feel that their attitudes are awry, that they do not subjectively experience their heteronomy, seems only to compound, not cancel out, the problem.

But subjectivist accounts implicitly recommend such delusional states as a cure for heteronomy. For one way to cure a person's heteronomy, on such views, would be to manipulate their higher order desires/values etc, so that those corresponded with the hitherto alienated first-order desires. If heteronomy is merely subjective in that subjectively felt alienation is constitutive or essential to it, one could remedy heteronomy by manipulating the higher-order desires/values etc, whose conflict with those first-order desires makes for their alienation. On this view, psychosis could be a remedy for neurotic heteronomy: my alienation from my neurotic desires to perform what I see as irrational actions can be eliminated if I come to see them – falsely – as rational and valuable actions, as things worth wanting to do. A philosophical elucidation of the concept of autonomy that entails that you can avoid violating someone's autonomy by manipulating them more, or that you can escape the heteronomy of neurosis by regressing to psychosis, is hardly promising.

In response, others, such as John Christman, offer procedural qualifications, so that an agent is not autonomous if their effective desires are those, which,

[44] H Frankfurt, 'Freedom of the Will and the Concept of a Person' in H Frankfurt (ed), *The Importance of What we Care About* (Cambridge University Press, 1988).

[45] G Watson, 'Free Agency' in G Watson (ed), *Agency and Answerability: Selected Essays* (Oxford University Press, 2004).

[46] D Velleman, 'What Happens When Someone Acts?' (1992) 101 *Mind* 461.

[47] M Bratman, 'Reflection, Planning and Temporally Extended Agency' in M Bratman (ed), *Structures of Agency* (Oxford University Press, 2007).

[48] As in Frankfurt-style examples.

even if they now identify with them, they would reject if they understood their genesis.[49] This is designed to ensure that desires brought about by manipulation or pathological irrationality are not counted as autonomous, regardless of whether or not the agent feels alienated from them. But it ultimately remains a form of coherentist subjectivism that runs into analogous problems in harder cases. The hypothetical conditional that, of a heteronomous desire (or belief, emotion, etc), the agent would not endorse it if they understood its causal history, only holds if the agent endorses appropriate procedural norms, such that they could recognise the normative force of procedural violations in such a way that they would be led to self-criticism by procedural scrutiny of their attitudes.[50]

Whether manipulation undermines someone's autonomy will thus depend on that person's normative conception of appropriate versus manipulative procedures, that is, on that person's conceptualisation of autonomy. One could thus avoid manipulating someone simply by ensuring that in addition to whatever else in their motivational set you manipulated, you also manipulated their conception of autonomy, their normative understanding of the distinction between the procedurally appropriate and the manipulative. We could cure someone of their heteronomy by manipulating their commitments as to the appropriate procedural requirements for desire-formation or as to the relevance of procedural standards themselves to the question of whether they should endorse a desire. Similarly, psychosis could still be a remedy for neurosis as long as one's delusions included delusions about the nature of and procedural requirements for autonomy. The more sophisticated subjectivist accounts of autonomy and heteronomy become, it seems, the more insidious become the sorts of manipulation they recommend.

Of course, subjectivist accounts of autonomy articulate something, which in many cases is valuable (internal harmony in a person's motivational set), but they cannot recognise that it is not always, in the abstract, valuable, nor explain why it is when it is. What such accounts lack is the ability to distinguish between those cases where the satisfaction of subjectivist criteria is a genuine mark of autonomy, and those cases where it may be a mark of greater heteronomy.

Now Pickard has elsewhere claimed that agency does entail autonomy – that spontaneity powers entail autonomy powers, so that choice, control and knowledge regarding one's body's particular movements entails choice, control and knowledge regarding one's ends as well.[51] But her argument there really shows only that on a certain dominant theory of action heteronomous action is incoherent. The theory of action (and thus the metaphysics) presupposed

[49] See J Christman, 'Autonomy and Personal History' (1991) 21 *Canadian Journal of Philosophy* 1; and the critical discussion in F Freyenhagen and T O'Shea, 'Hidden Substance' (2013) 9 *International Journal of Law in Context* 53.

[50] See Freyenhagen and O'Shea, ibid 63.

[51] H Pickard, 'Psychopathology and the Ability to Do Otherwise' (2015) 90 *Philosophy and Phenomenological Research* 135.

by subjectivist accounts – the so-called 'standard story'[52] – is on independent grounds inadequate. The standard story conceives of action as bodily movement that is just caused by desire (plus belief), and heteronomous actions as bodily movements that are just caused – by alienated desires, desires that are in conflict with the higher-order desires (values, etc) that supposedly 'speak for' the agent. Since it is, in this picture, the higher-order desires that represent the agent's role in their action that a heteronomous action is caused without and contrary to the agent's higher-order desires means that it occurs without the agent's own participation. But we cannot make sense of heteronomous agency by appeal to the idea that such actions are just caused without the agent's participation, for then it is unclear how those 'actions' are really actions at all.[53]

But it does not follow from that that agency entails autonomy, that the concept of heteronomous action is incoherent. What Pickard's account misses is that the standard story of action with its event-causal metaphysical presuppositions does just as bad a job of making sense of autonomous as of heteronomous action.[54] If the standard story cannot make sense of autonomous action anyway, its failure to make sense of heteronomous action says nothing against the reality of heteronomous action itself. The proper conclusion to draw is not that agency is necessarily autonomous, but that the standard story is inadequate to the task of rendering action, autonomous or heteronomous, intelligible.

Heteronomous action must consist in an alien structure in the agent's relation to the ends they themselves realise in their actions. But this cannot be a merely subjective affair; since subjectivist accounts are systematically unable to articulate what is valuable in subjective autonomy, only some form of objectivist account will do. Christman's position contains an element that is genuinely objectivist, in his stipulation that the process of desire-formation must not have involved 'manifestly inconsistent' desires or beliefs,[55] objectivist because it does not depend on whether or not the agent subjectively recognises the relevance of manifest inconsistency. But why, once we accept that something objective like this can undermine autonomy, should we stop there?

A robustly objectivist account of autonomy would make autonomy track responsiveness to real reasons.[56] Autonomy then refers to an ideal wherein the

[52] Velleman (n 46).

[53] M Alvarez, 'Actions, Thought-experiments and the Ability to Do Otherwise' (2009) 87 *Australasian Journal of Philosophy* 61; H Steward, *A Metaphysics for Freedom* (Oxford University Press, 2012).

[54] For a recent, persuasive argument to this effect see A Hinshelwood, 'The Relations Between Agency, Identification and Alienation' (2013) 16 *Philosophical Explorations* 243.

[55] Christman (n 49) 15.

[56] See S Wolf, 'Sanity and the Metaphysics of Responsibility' in F Schoeman (ed), *Responsibility, Character, and the Emotions* (Cambridge University Press, 1987); P Benson, 'Feminist Intuitions and the Normative Substance of Autonomy' in JS Taylor (ed), *Personal Autonomy* (Cambridge University Press, 2005). Freyenhagen and O'Shea (n 49) and F Freyenhagen, 'Autonomy's Substance' (2015) 34 *Journal of Applied Philosophy* 114 – to which the last few paragraphs are heavily indebted – recently defend the objectivist view.

powers of practically rational agency are properly developed and engaged. In autonomous action practically self-conscious thought is efficacious:[57] in acting, the agent knows the end they are striving to realise 'under the aspect of the end',[58] that is, they apprehend the end they are realising first-personally or from the inside, *qua* end, viz as something valuable, good, worth realising, and they apprehend what they are doing as issuing from their apprehension of the end. Hence the autonomous agent is in a position to answer Anscombe's distinctive sense of the question 'why?' by giving their reasons for action, which they apprehend, as such from the inside.[59] When this is so, the agent's relevant motivational states are 'transparent'[60] but for the right reasons: they are an expression of the agent's rational powers of responsiveness to the world. This conception underpins libertarian accounts of a Kantian or Sartrean character which claims that agency is necessarily autonomous, a view Pickard seems to endorse. But such libertarian accounts assert the ideal as an actuality – they assert the necessary practical efficacy of reason. Yet reflection on what we know about the real psychology of persons suggests otherwise – that though autonomous action, practically self-conscious thought, is a real possibility, it is not a necessary feature of agency. It is rather an aspiration and an accomplishment that can fail or break down because autonomy-powers can be prevented from developing or engaging properly.

V. THE MORAL PSYCHOLOGY OF HETERONOMY

What is at stake in thinking about autonomy and heteronomy is, to put it in Aristotelian terms, the possibility of a harmonious relation between the rational and non-rational natures in the soul.[61] And psychoanalysis, as Lear has argued, is a taking up of this Aristotelian question that rejects the picture of reason and nature sitting alongside one another as unrelated principles in the soul, and reconceives their relation as one of interpenetration.[62] This move allows it not only to further our understanding of how autonomy might in practice be realised, but deepens our appreciation of the ways it can go awry, and thus of why it should be a difficult accomplishment at all. Let me consider two such ways.

First, in subjective unfreedom/heteronomy, the end the agent realises in action is not one they apprehend 'under the aspect of the end': although it is moving

[57] See M Boyle and D Lavin, 'Goodness and Desire' in S Tenenbaum (ed), *Desire, Practical Reason, and the Good* (Oxford University Press, 2010); M Haase, 'Practically Self-Conscious Life' (MS); S Rodl, *Self-consciousness* (Harvard University Press, 2007).

[58] T Aquinas, *Summa Theologiae* La, question 6, art 2.

[59] See E Anscombe, *Intention* (Harvard University Press, 2000).

[60] In Moran's sense in Moran (n 40).

[61] See J Lear, *Wisdom Won From Illness* (Harvard University Press, 2017) ch 1.

[62] This is not something that cognitive psychology has been able to do, though I cannot defend this claim here.

them to action, and so they know that it is their end, they cannot make sense of it as their end. They see it as not sufficiently worth pursuing, but this evaluation does not have the result of dislodging or dissipating its power, its draw. The person's evaluative reason has become, with respect to a certain region of inner life, inert, because that region has been insulated from reason's power by a forcefield of unintelligibility. And this means reason has ceased, within this region of their life, to be practically efficacious. The agent knows, as if third-personally, what the end pursued in their action as a matter of fact is, but they cannot avow it as their end. It seems to them alien, and their alienation manifests a crisis of self-intelligibility. Such a structure has often been thought characteristic of neurotics and unwilling addicts 'helplessly violated' by their desires,[63] and of other even more troubling cases. Wollheim recounts the serial murderer Dennis Nilsen's thinking to himself, as he prepared to execute his latest victim, 'here I go again', and remarks that here 'he is thereby revealing … [that] His murderousness surprises him as much as us'.[64] Though 'we are very unlikely to be able to see why … he desires what he desires … [he] is likely to be in exactly the same position'.[65]

This is the sort of heteronomy that subjectivist accounts recognise, but as we have seen they are not able to explain why it is heteronomous, nor how such heteronomous action is even possible. But a certain sort of psychoanalytic account is promising on both scores, because it brings into view how desire may be organised by our non-rational nature – unconscious phantasy – through unconscious symbolic links, so that our rational powers of evaluation and practical deliberation are rendered impotent. As Lear has argued, Freud's conception of unconscious phantasy replaces the canonical conception of a simply irrational part of the soul with a non-rational form of mental functioning that is not simply chaotic but has an inner logos of its own, though a fundamentally different, archaic and infantile one which is metaphorical, wishful and bodily.[66] Phantasies can be thought of as emotionally and affectively charged implicit narrative or dramatic structures that organise conscious experience like an emotional *a priori* structure. Through symbolic associations and substitutions, unrealistic unconscious phantasy can get overlaid onto the world in distorting ways, such that the person is consciously motivated to pursue some end that is incomprehensible to them as an end. They cannot see any point or worth in it, and that is because, as a conscious end in the world, there is no point to it. It is really just a metaphorical stand-in for some archaic relic, an ossified infantile wish that can make no sense to adult, conscious comprehension, or is so unacceptable to their conscious reflection that they cannot get it into view.

[63] Frankfurt (n 44).
[64] Wollheim, *The Mind and Its Depths* (n 2) 124.
[65] ibid.
[66] See Lear (n 61).

Thus, phantasy can give rise to a desire that is 'imperious'[67] because it is unintelligible, and so untestable and unrevisable: acting on it does not promise any comprehensible satisfaction, not only as we see it, but as the agent themselves sees it; it becomes an entrenched bit of unreason. Such imperious desires are manifestations not of rational powers, but of the subversion of those powers by unconscious phantasy. This is why they lack transparency. They are not non-transparent in quite the way Moran seems to envisage the paradigm case, but although the agent first-personally apprehends their imperious desire in the distinctive way in which desire is felt from the inside, and thus apprehends from the inside the end as the end they are pursuing, they are not able to first-personally apprehend that end as an end, 'under the aspect of' the end': it is not intelligible to them as something worth pursuing. It is because the distorting influence of unconscious phantasy, rather than the agent's rational powers, is the source of the desire that its 'force' is so divorced from its 'importance'.[68] An agent's acting on an unimportant and unintelligible, alien desire appears inexplicable only until we grasp that what is explanatorily crucial is not the desire they are acting on, but the phantasy the agent is 'acting out'. Because they cannot apprehend nor comprehend this unconscious phantasy first-personally, nor the fact of their acting it out, they are not in a position to subject either to practical reflection, revision or restraint. If this is right, Freud's concepts of unconscious phantasy and of 'acting out' help elucidate subjective unfreedom as a real psychological possibility for us.

Second, objective unfreedom or heteronomy. The action is in pursuit of an end that the person apprehends from the inside not only as their end but as an end. The desire, which motivates them is thus, at the time of acting, transparent in Moran's sense: the person comprehends their desire from the inside and takes it to be tracking what is worth desiring. There is, at the time, no first-personal gap between force and importance. Objectively heteronomous action is thus subjectively, phenomenologically, indistinguishable from autonomous action. What distinguishes it from autonomous action is what it is that explains the transparency of the desire. In autonomous action the desire is transparent for the right reasons: it is transparent because it is the expression of the person's rational powers in shaping their own inclinations in accordance with their appreciation of what is worth wanting. Objectively unfree actions are motivated by desires that are transparent for the wrong reasons: they are rooted in unconscious phantasy, like imperious desires, but that phantasy has reorganised – distorted – not only their desires but also their view of the world and of themselves, their view of external and internal reality.

The direction of fit is the wrong way round here: desire has not been brought into line with reason; reason – including perceptual and introspective

[67] Wollheim, *The Thread of Life* (n 38) 120.
[68] See Wollheim, *The Mind and Its Depths* (n 2) 123.

interpretation – has been brought into line with desire, because both have been conditioned non-rationally by unconscious phantasy. There is, then, no immediately available subjective foothold by which the person can get into view either sort of distortion. If such distortion is, unusually, deep and stable, we call it brainwashing or psychosis, but more localised and transient distortions of this kind are commonplace, figuring centrally in some types of personality disorder, and more commonly still at lower levels in everyone.

Subjective unfreedom is thus an advance on objective unfreedom, because it manifests the person's capacity to avoid distorting the world to fit the desire, which, with the world undistortedly in view, is encountered as simply unintelligible instead.[69] Hence the arachnophobe's inefficacious awareness that British spiders are harmless contrasts favourably with psychotic delusion that, say, a cover-up is suppressing the facts about the actual deadliness of British spiders. More familiarly, anxiety, anger or affection may distort our experience of persons and situations rather than being shaped by those experiences, and when this happens our experience is prefabricated by unconscious phantasy rather than being a rationally responsive apprehension of the world.

The power to distinguish what is coming from inside and what is coming from outside is what Bion called the capacity to contain bits of mental life,[70] and it is something that must be acquired through the right intersubjective experiences, with, to begin with, the parents. The more developed the capacity to contain, the more a person will be able to avoid unconscious distortions of the world in experience. Uncontained phantasy will often be localised, transient or oscillating: the person may think to themselves that their earlier anger was quite irrational given the context but be unable to retain this thought first-personally the next time it happens. Outside the immediate grip of the phantasy, they apprehend the temporary transparency of those motives as itself an illusion, but in the moment such insight is unavailable. Phantasies may be activated by events and then pass away so that temporary distortions of experience wax and wane.

Through the concepts of unconscious phantasy and of containment, then, psychoanalysis also helps elucidate the possibilities of mundane objectively unfree action.[71]

VI. THE ANTINOMY OF RESPONSIBILITY

This account of heteronomy elucidates the possibility of unfree agency, which explains the pervasiveness of the tyranny of the past that is, I have claimed,

[69] See Harcourt (n 40).

[70] W Bion, 'Attacks on Linking' in W Bion (ed), *Second Thoughts* (Karnac, 1984).

[71] For readers wary of psychoanalytic vocabulary, the points I have just made could be roughly articulated in the terms of the contemporary cognitive-neuroscientific paradigm – though not without loss.

the true object of reflection on the whole reality of the person. We now have an explanation of why attention to the whole reality of the person rationally undermines affective blame, objectively, such as to undermine blameworthiness and thus the normative structure of our holding criminally responsible practice itself.

The upshot, I want to claim, is that our practice of holding responsible is fundamentally inadequate to the real psychology of criminal law's persons. The task for the philosophy of criminal law is to engage in concrete utopian reflections on the possibility of a different responsibility practice in a changed form of ethical life that would be more adequate to the real psychology of persons.

This is not, of course, the conclusion Lacey and Pickard reach. Instead, we saw, they insist that the sensitivity of our moral emotions to the whole reality of the person has no bearing on the validity of the normative structure of our practice of holding criminally responsible, 'detached blame'. But Lacey and Pickard individually have very different conceptions of what sort of thing our practice of holding responsible is, and together they seem to follow Pickard's view of responsibility norms as embodying timeless metaphysical truths about agency, just the kind of view which Lacey has elsewhere criticised.[72] But this means they have to assert that our existing responsibility norms track capacities that are properly exercised in wrongdoing even where affective blame is not apt. And this overlooks the distinction, considered above, between agency simpliciter, spontaneity and autonomous agency. It ignores the actuality of the category of unfree agency and the real moral psychology of heteronomy.

Elsewhere, Lacey has argued that privations in rational capacities undermine autonomy in ways the criminal law's norms fail properly to recognise, and in ways she thinks deprive many offenders of a fair opportunity to conform their behaviour to the criminal law.[73] But these same factors become, in the Lacey and Pickard position, irrelevant to the norms of holding criminally responsible. Lacey's independent arguments suggest that the Lacey and Pickard account of detached blame – that 'responsibility tracks agency' – is inadequate. She has railed against retributivists like Moore for advancing the view that responsibility tracks metaphysical truths about agency, but the 'responsibility without blame' model seems to endorse just that sort of view of the norms of our practice of holding criminally responsible.

Moreover, the 'responsibility tracks agency' view is, as they admit, a rather simplified story, for it says nothing about how the central excuses such as duress operate, but it is clear that duress does not rest on the fact that the person's agency was not engaged, that their actions were not fully actions; duress

[72] See N Lacey, 'Responsibility and Modernity in Criminal Law' (2001) 9 *Journal of Political Philosophy* 249; N Lacey, 'In Search of the Responsible Subject' (2001) 64 *Modern Law Review* 350.

[73] N Lacey, 'Socializing the Subject of Criminal Law: Criminal Responsibility and the Purposes of Criminalization' (2016) 99 *Marquette Law Review* 541.

excuses by negating the voluntariness of action, not of movement. If we are to understand our current conception of the bounds of voluntariness as being more than merely decisionistic, we will need to see it as recognising – albeit in certain very limited circumstances – the fragility and contingency of autonomy, the vulnerability of our rational powers to inner and outer circumstance: the real psychology of persons, and the social context in which that develops and in which people act. But then the 'responsibility tracks agency' view must be wrong. Responsibility norms already implicitly recognise (albeit in limited ways) both that responsibility depends on autonomy, and that agents are not necessarily autonomous.

Lacey and Pickard resist this conclusion because, I suspect, they worry that the only alternative to holding responsible is not holding responsible, exculpating. That alternative is unattractive because to simply exculpate heteronomous wrongdoing would be ethically unserious. This is already a difficulty for Lacey's account in 'Socializing the Subject', for she there also links her 'socializing' of the criminally responsible subject to a vision of 'criminal law and its surrounding processes' as 'aspir[ing] to foster positive goals such as integration, reform, and even forgiveness',[74] but it is unclear how these could be pursued if responsibility norms exculpate heteronomous wrongdoing in the ways she envisages (whether through more contextual *mens rea* standards or defences).[75] Within the binary logic of our practice of holding criminally responsible, holding not responsible, exculpating marks the end of the matter. And that would surely fail to do justice to the person's relationship to their past wrongdoing and to the community's relationship to that and to the person in light of it.[76]

Both the Lacey and Pickard position which asserts that 'responsibility tracks agency' and that our practice of holding responsible is unscathed by the critique of affective blame, and the alternative that Lacey endorses in 'Socializing the Subject', which entertains not holding responsible in recognition of individual heteronomy, fail to carve at the joints of moral reality. What motivates each is the inadequacy of its opposite. Neither holding responsible nor exculpating is satisfactory, and yet there seems to be no other alternative: this I call the *antinomy of responsibility*.

What are we to make of it? Lacey and Pickard side with holding responsible followed by forgiving and reparative dialogue, and that is certainly a more attractive compromise than we have in practice at the moment, but a compromise it still is, and an unstable one. This becomes explicit when Lacey and Pickard

[74] ibid 556.

[75] Lacey prefers more sensitive fault standards, while others have favoured special defences: D Delgado, 'Rotten Social Background Should the Criminal Law Recognise a Defence of Severe Environmental Deprivation?' (1985) 3 *Law and Inequality* 9.

[76] The alternative of a 'bar to trial' defence (see Duff, *Answering for Crime* (n 25) 192–93; Duff, *Punishment, Communication, and Community* (n 9) ch 5) would have the same result and seems no less ethically unserious.

concede that, as they are seeing things, holding responsible sits in tension with reparative dialogue and change.[77]

The categorial straitjacket that forces us into the antinomy of holding responsible or not holding responsible is the artefact of a concrete, contingent form of ethical life. The alternative to compromise would be to reject the choice our existing responsibility practices impose on us: those are inadequate to the moral reality that our moral emotions intimate and a realistic philosophical psychology elucidates. If we are willing to put that practice in question – to see it as merely the form of ethical life that we happen to have – then we can see that not holding responsible does not entail exculpating after all: it only entails exculpating within the categorial terms of our existing practice. A radically different responsibility practice might make possible a response, which is neither holding responsible nor exculpating.

The question, then, is whether a different form of ethical life is possible, whether a different responsibility practice informed by the real psychology of persons that could recognise heteronomy and institute realistic respect for unfree agents is possible. I see no convincing grounds for maintaining that our ethical life is in principle the most desirable one possible, that it could not conceivably change to take on a very different, more realistic and truthful shape, or for insisting on a pragmatism, whether Wittgensteinian or Hegelian, that renders the idea of radical ethical criticism of form of ethical life itself unintelligible. That way lies, as Williams put it, 'an undiscriminating acceptance of whatever conceptual resources of [our] society actually exist'.[78] Our efforts should be turned to the task of thinking through what such a different responsibility practice, a more realistic and truthful form of ethical life, might look like and what resources for its development might be identified in what we already know about the as-yet only partially actualised human life form.[79]

VII. RESPONSIBILITY, REIFICATION AND RESPECT

I began by noting the influential *reification objection*: retributivists have often suggested that any alternative to our existing practice of holding responsible must reify agents, treating them as if they were a mere thing and thus depriving them of the respect that is owed to persons. As Strawson put it, 'the humanity of the offender himself is offended' by not holding responsible.[80] Arguments in these terms have been commonplace since: not to hold responsible would be

[77] Lacey and Pickard, 'To Blame or to Forgive?' (n 4) 690.

[78] B Williams, 'Pluralism, Community and Left Wittgensteinianism' in B Williams (ed), *In the Beginning Was the Deed* (Princeton University Press, 2005) 36.

[79] See C Reeves, 'Beyond the Postmetaphysical Turn' (2016) 16 *Journal of Critical Realism* 217.

[80] Strawson (n 5) 62.

to 'objectify',[81] 'disrespect'[82] or 'dehumanise'[83] the person, to deprive them of the status of persons. Channelling the 1970s anti-psychiatry zeitgeist, Murphy wrote:

> Practices of punishment and responsibility are compatible with human dignity in that they place a premium upon the status of persons as choosing beings. One alternative to this is coercive therapy … [involving] perhaps a total restructuring of the personality [as in] *A Clockwork Orange* [or] *One Flew Over the Cuckoo's Nest*.[84]

Murphy echoes a contrast pressed earlier by Strawson,[85] and, though presented as an alternative, its rhetorical force flows from the tacit suggestion that a reifying system of treatment in which crime is 'regarded as a symptom', a 'happening with a causal explanation rather than an action for which there were reasons',[86] is the *only conceivable* alternative to our existing practice of holding responsible. Any practice that departed fundamentally from our practice of holding responsible would be one in which the 'distinction between mere events or happenings and human actions is erased',[87] reifying people and thus failing to respect them as agents or persons with dignity and rights.

Now this reificatory denial of respect is, so the objection goes, entailed by the acknowledgment of heteronomy. Once we admit that someone acted heteronomously, we are effectively admitting that they are not agents or persons at all, at least in respect of the relevant conduct. And once that move is made, the individual's rights, their entitlement to be treated as a person, have been obliterated, for as a bit of mere causal nature, there is no reason not to coercively treat them, as we might a dangerous animal. Any alternative to holding responsible that countenances the heteronomy of criminal law's persons is bound to regard someone, as Moore puts it, 'as an in-itself rather than as a for-itself'.[88]

Yet the objection rests on an inadequate conception of the object of respect, real persons, and in turn on an austere and unrealistic conception of respect for persons. It is rooted, as Moore's Sartrean jargon indicates, in the Kantian dualism of things and persons, where heteronomy belongs to the realm of things, mere causal nature, whereas respect attaches to persons conceived as necessarily autonomous, rational wills.[89] With that dualism in place, to view someone as heteronomous entails a shift to seeing them as a mere thing, because the hallmark of personhood is rational autonomy.

[81] Moore (n 12) 546.
[82] Tadros (n 13) 392.
[83] Gardner (n 14) 254.
[84] Murphy (n 24) 134–35.
[85] Strawson (n 5).
[86] H Morris, 'Persons and Punishment' in H Morris (ed), *On Guilt and Innocence* (University of California Press, 1976) 36.
[87] ibid 37.
[88] Moore (n 12) 546.
[89] See I Kant, *Groundwork of the Metaphysics of Morals* (trans H Paton, Harper and Row, 1964) 65 and 428. I discuss this point in more detail in C Reeves, 'Adorno, Freedom and Criminal Law' (2016) 27 *Law and Critique* 323.

Retributivism accurately registers the fact that our existing responsibility practices are deeply embedded in this dualism, so that within its conceptual constraints, to not hold responsible is to reify. But this implication arises only because of the way in which our responsibility practices and their underlying conceptual order shoehorn the ethical phenomena into a false dilemma between autonomous persons and heteronomous things. Since this Kantian picture rules out heteronomous agency, it confines respect to autonomous agents. The question of the normative standing of heteronomous, unfree agents simply does not arise. And under the influence of this picture, it is supposed that there is no alternative to our practice of holding responsible that would not be reifying.

But if agency can be unfree, heteronomous, in the ways I have suggested, then the picture is false; it ought to be abandoned. The forms of heteronomy I have considered do not undermine agency: recognising them does not imply a view of actions as indistinguishable from mere happenings, symptoms; they do not turn someone into a mere thing or mere animal, a bit of causal nature. Rather, they work through agency and action: heteronomy is a privation in the exercise of agency, not a negation of it. Heteronomy does not obliterate agency; it is intelligibly predicated only of agency. And this means there should be no temptation to insist that heteronomous agents are mere things that do not warrant respect.

Indeed, if we are all, to a greater or lesser extent and in differentiated and particular ways, heteronomous agents, heteronomous agency should be seen as the norm and the paradigm case of a rights-bearing agent. Moreover, autonomy and heteronomy are not all-or-nothing. Heteronomy is typically localised, selective and specific. Regions of heteronomy in a person's life, rather than heteronomy across the board, are the norm. Such regions do not render a person wholly unintelligible to us as an ethical interlocutor. Rather, they render certain regions of that person's life – their desires, beliefs, emotional responses, in some areas – unintelligible to us: just the same regions that are (in subjective heteronomy) or should be (in objective heteronomy) unintelligible to them.

Consider, then, Gardner and Macklem's claim that 'self-respecting defendants' have reason to refuse excuses based on 'rationally incapacitating conditions' because they

> have an interest in being accorded their status as fully-fledged human beings, ie as creatures whose lives are rationally intelligible even when they go off the rails, and who can therefore give a rationally intelligible account of how they came to do so.[90]

This implicitly denies the possibility of heteronomy. As MacIntyre writes,

> individuals afflicted by neurosis resort to psychoanalysis [because] they have found themselves doing things that they have no good reason to do or good reason not to

[90] J Gardner and T Macklem, 'Compassion Without Respect? Nine Fallacies in *R v Smith*' (2001) *Criminal Law Review* 623, 627.

> do, they have tried to reason with themselves and then have discovered that their reasoning has been flawed by phantasy, or … has been practically ineffective. They have become to some degree unintelligible to themselves.[91]

Certainly, we want to be accorded 'our' status as 'fully-fledged human beings', and to be such is to aspire to being a creature whose life is 'rationally intelligible'. But this is as a matter of fact frequently not what our lives are actually like: we are endemically liable to fall short of that aspiration. But this does not mean we cease thereby to be fully-fledged human beings, becoming non-rational animals, or, even, mere things, our actions mere happenings. It is internal to being a fully-fledged human being that one is a creature whose life is not necessarily and always rationally intelligible.

In cases of objective unfreedom, the person is, at the time, unable to see their own heteronomy. Under such conditions, an agent's seemingly 'rationally intelligible account' of their conduct may be false, and systematically so. The drive to be the sort of creature that can give a rational self-account may, under conditions of objective heteronomy, feed mere rationalisation and thus entrench objective heteronomy. After all, the tendency to render intelligible what is in fact unintelligible lies at the heart of objective heteronomy itself, and surely our interest is in being able to give a genuine rationally intelligible account of ourselves rather than a mere rationalisation.

With subjective unfreedom, the person is aware of their unfreedom, because they are aware of the objective unintelligibility of their alien end. Here, the agent will not be able to give a 'rationally intelligible account' of themselves at all. Relative to objective heteronomy this is an advance, but it still manifests a privation *qua* rational animal. Yet, such self-unintelligibility is not a failure to be fully a person; it is a possibility of privation immanent to personhood.

Unfree agency occupies a space between autonomous self-intelligible agency and arbitrary causal nature, where transparent, rational self-intelligibility is an essential aspiration, but not a given. The reification objection is premised on an austere and unrealistic psychological and metaphysical picture that is unable to make space for this thought. In the equation of 'fully-fledged human being' with 'rationally self-intelligible creature', it, like our practice of holding responsible, is guilty of over-rationalising the psychology of real persons. This over-rationalising of the person goes hand-in-hand with an under-rationalising of heteronomy: recall that unconscious phantasy is not remotely mere causal nature, but rather an idiosyncratically meaningful drama with a certain narrative and bodily logos of its own.

The conceptual dualism of our holding responsible practices tracks the organising opposition of persons and things and forces us to fit the facts into one of these categories. But it is wrong to think that this dilemma is itself basic

[91] MacIntyre (n 39) 13.

and unavoidable rather than the artefact of the moral practices we happen to have. Lacey and Pickard wrongly concede that the only way to institute respect is by holding responsible, and this obscures the more radical implications of their thought that reflection on the whole reality of the person (rationally) undermines affective blame. They endorse our existing holding responsible practice because they see that as the only alternative to reification. But this dilemma itself presupposes the categorical frame of our existing responsibility practice – the modern metaphysics of things and persons – and the austere conception of what persons would have to be like to warrant the respect that that entails. Lacey and Pickard's position really points to the need for radically changed, psychologically realistic and ethically serious responsibility practices instituting real respect for persons comprehended as potentially free but often unfree agents.

VIII. THE GRAMMAR OF TAKING RESPONSIBILITY

One promising avenue is suggested by Lacey and Pickard's remark that 'it is a presumption of effective treatment that patients have choice and a significant degree of control over their behaviour and can therefore be asked to take responsibility for it, as we naturally say'.[92] Now asking someone to take responsibility is not the same as holding them responsible. And the grammar of taking responsibility may, it seems to me, contain the conceptual seeds for a changed responsibility practice that would be truthful and ethically serious, that could institute genuine respect for persons apprehended as rational animals for whom privation and heteronomy are pervasive possibilities, and that could be more in tune than our existing practice is with the reparative and reconciliatory aspirations that animate Lacey and Pickard's account.

That grammar of taking responsibility is at the core of therapeutic action – not the cognitive therapy on which Pickard focuses, but the psychoanalytic conversation, in which 'two people actively interact with each other, each in the process (among other things) of trying to understand each other and themselves'.[93] In asking someone to take responsibility within the psychoanalytic process, the analyst aims to cultivate in someone self-intelligibility-for-the-right-reasons, which is to say, rational health, real autonomy, but this necessarily involves their transforming their relationship to their own mental life and to what they have been and have done. This undertaking necessarily involves a commitment to openness to truthfulness, to the reality of others, and to potential change. It builds into its presuppositions humility in the face of one's vulnerability as well as the aspiration to and courage for a painful transformative self-emancipatory

[92] Lacey and Pickard, 'From the Consulting Room to the Courtroom?' (n 4) 13.
[93] Lear (n 61) 179.

process. That is, it offers a more psychologically realistic and ethically serious answer to the question, in Lear's words, of what it would involve for someone to become 'able to reconcile themselves to their past'.[94]

Whereas the grammar of holding responsible is that of theoretical attribution of a status, a being, and is essentially second-personal, the grammar of taking responsibility is that of a dynamic, practical process of becoming, and is essentially first-personal. So while holding oneself responsible is essentially an internalised second-personal theoretical judgement in which the self is a passive object, taking responsibility is an active practical judgement in which the 'I' figures as an agent of a self-transformation. Holding someone responsible is to ask them to hold themselves responsible, to internalise a second-personal static, theoretical judgement; asking someone to take responsibility is to ask them to undertake a process of change towards truthfulness and increased freedom.

The grammar of taking responsibility thus presupposes neither that someone is a necessarily heteronomous, passive mere thing, which would make emancipatory change impossible, nor that they are a necessarily autonomous person, which would make emancipatory change unnecessary. Rather, it presupposes that they are a realistically conceived 'fully-fledged human being', that is, a potentially autonomous but actually (partially) heteronomous person, one whose unfree passivities are themselves possible objects of their own self-emancipatory agentive activity.

On this view, it seems to me, Lacey and Pickard's rejection of affective blame – which I have so far accepted – would have to be revisited. For within the framework of taking responsibility, rather than holding responsible, the possibility opens up for a different kind of 'affective blame' – a different kind of anger – that is distinct from the kind that is bound up with holding responsible. It would be a kind of anger that is tied to the recognition of someone as a complex, messy whole person in a complex, messy context, quite opposed to the sort of anger that is involved in the affective blame that Lacey and Pickard rail against. Such anger, insofar as it is embedded in a practical relation to the other as a contextualised, complex whole person might be not only acceptable but required.

A premise of psychoanalysis is that a person needs the right sort of conversation with another in order to take responsibility in this way. If the psychoanalytic conversation is to help the person take responsibility, the analyst must be able to engage truthfully, that is, they must be able to take responsibility themselves as well. The grammar of taking responsibility, then, is genuinely dialogical, governing a shared intersubjective undertaking that presupposes humility and courage and a willingness to be truthful and an openness to change

[94] Lacey and Pickard, 'To Blame or to Forgive?' (n 4) 690.

on both sides. If the grammar of taking responsibility were to be instituted practically in a form of ethical life, it would require that in asking someone to take responsibility the community also be willing to take responsibility, that is, to engage truthfully with what it has been doing and to seriously undertake to change. Needless to say, this implies a level of political maturity that seems to be beyond the modern *polis*.