Philosophica 85 (2012) pp. 11-33

ON THE COMPATIBILIST ORIGINATION OF MORAL RESPONSIBILITY

Stefaan E. Cuypers[*]

ABSTRACT

Derk Pereboom defends a successor view to hard determinism in the debate on free will and moral responsibility. Pereboom's hard incompatibilism challenges libertarians and compatibilists alike to address the problem of origination. In the present article, I discharge this task on behalf of compatibilism.

1. Introduction

Apart from general agency requirements — such as being capable of intentional action, rational deliberation and evaluative judgement — moral responsibility more specifically has epistemic, control, and authenticity requirements. Accordingly, an agent *S* cannot be morally responsible for a particular action *A* unless (1) *S* knows, or beliefs, that *S* is doing wrong (or right) in performing *A*, (2) *S* exercises responsibility-relevant control in doing *A*, and (3) *A* stems from psychological

[*] I am indebted to two anonymous referees for *Philosophica* for their critical remarks and fruitful suggestions, and I am very grateful to Tom Claes and Tim De Mey for organizing the Ghent Conference Moral Responsibility: Analytic Approaches, Substantive Accounts and Case Studies in October 2010.

antecedents, or “springs of action”, that are authentic. That there is *some* cognitive and *some* control condition for moral responsibility is not controversial.[1] However, whether the authenticity, “source”, or “origination” condition is a bona fide requirement for moral responsibility is controversial.[2] The central idea behind the last of these conditions is that the agent has to be the ultimate source of the actions for which he is morally responsible; that he should be the real originator who controls the springs of his morally responsible action. The agent can

[1] As knowledge entails truth, some theorists (for example, Smith 1983) endorse the objective view that an agent is morally responsible for an action only if performing that action is objectively wrong (or right). Others (for example, Haji 1998, chap. 9) maintain the subjective view: a belief condition — the agent’s believing that performing the action is wrong (or right) — is sufficient to fulfill the epistemic requirement. Various accounts of the control or freedom condition have been proposed in the literature. Some theorists (for example, van Inwagen 1983) have argued that a person has the right sort of control only if he had genuine alternatives — he “could have done otherwise”. Others (for example, Frankfurt 1988) have maintained that a person has the pertinent kind of control just in case he identifies with the action’s motivating desires. Still others (for example, Fischer and Ravizza 1998) have suggested that a person has the required control — “guidance control” — only if he is appropriately sensitive to reasons; he would, under specified conditions, have arrived at some other decision were apt reasons present. And yet others (for example, Mele 1995) have defended the view that the germane control consists in the action’s being produced non-deviantly by causal antecedents such as desires, beliefs, values, and so forth that satisfy certain constraints.

[2] The authenticity (source, origination) condition is also called the “ultimate control” condition for moral responsibility. If one uses this latter terminology, then one should clearly distinguish the third condition from the second, by calling the second the “proximate” or “local control” condition. Other verbal equivalents of the third condition in the literature are the “authorship” or “autonomy” condition.

shoulder moral responsibility for actions only if they, and the decisions which produce them, are authentic or "the agent's own".

Recently, Derk Pereboom has exploited the authenticity condition to argue that we do not have the sort of free will required for moral responsibility. The pivotal plank in Pereboom's overall argument to establish his so-called "hard incompatibilist" position — a successor view to hard determinism[3] — is, what he calls, "the origination principle O". According to him, neither libertarian nor compatibilist accounts of free will can, conceptually or empirically, comply with this crucial principle. Pereboom's exacting position can be interpreted as a challenge for libertarians and compatibilists alike to propose a solution to the problem of origination.

In this paper, I take up Pereboom's challenge and offer a compatibilist answer to this central question about sourcehood. After presenting Pereboom's challenge (section 2), I develop a compatibilist solution to the origination problem (sections 3 and 4). In conclusion, I explicitly respond to the details of Pereboom's challenge in light of my "Forward-looking Relative Authenticity" account of ultimate origination (section 5).

2. Incompatibilist ultimate origination

To begin with, some rehearsal of the basic terminology will be helpful. Determinism is "the thesis that there is at any instant exactly one

[3] Pereboom's view cannot, strictly speaking, be called "hard determinist" in the classical sense. For that reason Robert Kane (2002, p. 27) calls views like that of Pereboom, Smilansky, G. Strawson and Honderich "Successor Views to classical hard determinism".

physically possible future." (van Inwagen 1983, p. 3) If this thesis is true, the facts of the past, together with the laws of nature, entail all facts of the present and future. Indeterminism is the denial of determinism. Compatibilism is the view that free will, free action and moral responsibility are compatible with determinism; incompatibilism is the denial of compatibilism. Libertarians are incompatibilists (and indeterminists) who believe that at least some of us, at times, perform free actions for which we are responsible. Hard determinists are incompatibilists who deny that we have the sort of free will required for moral responsibility; hard incompatibilists deny the same, irrespective of accepting determinism or not.

Pereboom (2001, pp. 2-6) distinguishes further between "leeway" and "causal history" incompatibilism. The former type of incompatibilism maintains that an agent's moral responsibility for an action depends primarily on the existence of alternative possibilities, commonly captured by the Principle of Alternative Possibilities (PAP): a person is morally responsible for what he has done only if he could have done otherwise.[4] The latter type holds that the most fundamental incompatibilist principle for explaining an agent's agential responsibility concerns not alternative possibilities but the actual causal history of an action: a person is morally responsible for what he has done only if he is the ultimate source of what he has done.

I agree with Pereboom's contention that of the two incompatibilist intuitions — the requirement of alternative possibilities and the requirement of sourcehood for moral responsibility — the latter is the deepest, most fundamental and plausible one. To be blameworthy or praiseworthy for an action, one has to be the ultimate source or cause of the action. For an agent to be morally responsible for an action, he must be its source in an especially strong sense. According to Pereboom, this core incompatibilist claim about origination can be expressed as follows:

[4] For discussion, see Widerker and McKenna 2003.

> *Origination Principle* (O): “If an agent is morally responsible for her deciding to perform an action, then the production of this decision must be something over which the agent has control, and an agent is not morally responsible for the decision if it is produced by a source over which she has no control.” (Pereboom 2001, pp. 4, 47, 54)

The second, negative part of O is also expressed by van Inwagen’s so-called “direct argument” for the incompatibility of moral responsibility and determinism (van Inwagen 1983, pp. 182-8):

If causal determinism is true, then there is some state of the world in the distant past *b* that is connected by the laws of nature to any action *A* that one performs in the present. But since no one (alive now) is even partly morally responsible for the state of the world *b* in the distant past, and no one is even partly morally responsible for the laws of nature that lead from *B* to *A*, it follows that no one is even partly morally responsible for any action *A* that is performed in the present.[5]

Since no one has control over the state of the world *b* in the distant past nor over the laws of nature, no one is morally responsible in a deterministic world. But also in an indeterministic world without agent-causation no one is morally responsible, because no one has control over decisions (and ensuing actions) if they are not produced by anything at all, if they occur without any cause whatsoever. Neither deterministically produced antecedents of action nor randomly indeterministically produced ones are events over which the agent has control. Accordingly, from the general principle O, the more specific principle of Pereboom’s causal history incompatibilism follows directly:

[5] This summary of the argument is from Fischer and Ravizza 1998, p. 153.

> *Causal History Principle* (CH): "An action is free in the sense required for moral responsibility only if the decision to perform it is not an alien-deterministic event, nor a truly random event, nor a partially random event." (Pereboom 2001, pp. 54, 89)

Pereboom explains the terminology of CH as follows:

> "We might call those events for which factors beyond the agent's control determine their occurrence *alien-deterministic events* and those that are not produced by anything at all *truly random events*. The range of events between these two extremes — for which factors beyond the agent's control contribute to their production but do not determine them, while there is nothing that supplements the contribution of these factors to produce the events — we might designate *partially random events*. By incompatibilist standards, an agent cannot be morally responsible for a decision if it is an event that lies anywhere on this continuum, because the agent does not have a suitable role in its production." (Pereboom 2001, p. 48)

In the light of principles O and CH, Pereboom argues that neither compatibilism nor an event-causal type of libertarianism can deliver an adequate account of ultimate origination — the sort of free will required for moral responsibility. Starting from the premise that, given principles O and CH, manipulation is responsibility-undermining, his line of argument in support of this conclusion goes like this (Pereboom 2002, p. 478).

1. Covert manipulation by a determining or randomizing manipulator undermines moral responsibility, since the victim has no control over the manipulator or external cause.

2. Causal determination presents no less of a threat to moral responsibility than does covert manipulation, since there is no relevant and principled distinction between an ordinary deterministic causal history and a manipulated one.[6]

3. Event-causal indeterministic histories are no less threatening to moral responsibility than deterministic histories, since there is no relevant and principled distinction between an ordinary indeterministic causal history and a manipulated one.

∴ Therefore, event-causal deterministic or indeterministic histories are responsibility-undermining.

According to Pereboom, only an agent-causal type of libertarianism can comply with principles O and CH, because agent-causes, as primitive active powers, are ultimate sources of morally responsible action. But although a conception of ourselves as morally responsible agent-causes is not incoherent (Pereboom 2004), there is little, if any, empirical evidence to believe that we are in fact agent-causes. Consequently, given that no position can secure, conceptually or empirically, the kind of free will required for moral responsibility, Pereboom's causal history incompatibilism radicalizes into *hard* incompatibilism: we live without free will and no one is ever morally responsible for anything he or she does.

Pereboom's hard incompatibilism challenges both compatibilists and libertarians — all believers in free will. One way to retort and to uphold one's favoured anti-sceptical position is by rejecting principle O (and CH). Accordingly, some compatibilists and some libertarians deny that sourcehood is a bona fide condition for moral responsibility (see, for example, Berofsky 2006). Alternatively, if one accepts O but shies away from its consequences, then one faces the problem of origination. Unless one is willing to adopt hard incompatibilism, one has then to discharge

[6] Pereboom (1995, pp. 245-9; 2001, pp. 110-117; 2007, pp. 93-8) bolsters up this premise by his so-called "four-case argument".

the burden of giving an adequate account of ultimate origination. Accepting O as a *compatibilist*, I take up the gauntlet. For a compatibilist, addressing the problem of origination amounts to tackling the second premise (2.) in Pereboom's argument for hard incompatibilism. I will argue that Pereboom's argument, although valid, is unsound because its second premise is false.

3. Compatibilist ultimate origination

Although a conception of ultimate origination or control is usually associated with libertarianism, also compatibilists can develop an account of such control that is compatible with determinism. Consider the following proposal. If an agent has ultimate control in performing an action (or taking a decision), then he is the ultimate originator or source of his action (decision). According to Ishtiyaque Haji (2009, p. 41), three conditions are sufficient for ultimate origination:

> "(i) The cause, or at least a causal antecedent, of the free action must be a component of the type of cause that plays a salient role in the production of *action* or *free action* (such as the having of a suitable belief or desire). The cause could not be something like the beating of an agent's heart. (ii) This cause (or part of it) must, in some obvious sense, be internal to its agent. (iii) The cause must be at least partly constitutive of the agent in a way in which, in virtue of being so constitutive, it would be correct to say that the action (or the free action) "truly" issues *from the agent*, or is the "*agent's own",* or is one over which *the agent has [ultimate] control*. It is something like (iii) that conceptions of ultimate origination seek to capture."

This account of ultimate origination is compatible with the supposition that any free action is deterministically caused. Libertarians then just have to add the fourth condition that the cause in (i), (ii), and (iii), must be a *non*-deterministic cause, or must itself be *non*-deterministically caused if it causes deterministically the free action. This extra condition takes care of non-deterministic or probabilistic causation as well as undetermined agent-causation.

Libertarians add a forth condition to the trio (i), (ii), and (iii), while compatibilists differ over what the correct interpretation of the third condition is. This condition gives a conceptualization of the agential control in principle O. Accordingly, Harry Frankfurt, for example, commonly regarded as a compatibilist, clearly accepts then a version of principle O, which I call the "Participation Principle":

> "A person is morally responsible for an action only if he is properly implicated (alternatively, "invested" or "engaged") in the action." (Frankfurt 1988)

David Velleman (1992, p. 470) crisply summarizes what the Participation Principle strives to encapsulate:

> "What primarily interests Frankfurt […] is the difference between cases in which a person 'participates' in the operation of his will and cases in which he becomes 'a helpless bystander to the forces that move him.' And this distinction just is that between cases in which the person does and does not contribute to the production of his behaviour."

According to Frankfurt, such participation involves the agent's having of second-order volitions — second-order desires concerning which first-order desires should move him to action. Such investment is a matter of activity on the agent's part that generates a set of first-order desires or attitudes he cares to have — desires internal to his “volitional structure” to which he decisively commits himself and with which he wholeheartedly identifies. The unwilling addict, who shoots up despite identifying with the desire to refrain from taking the drug, is not morally responsible for indulging. Although taking the drug is an intentional action on his part, the unwilling addict is not invested in this action; he does not participate in the operation of his will and is “passive” with respect to it.

In another place, I have argued in the context of the problem of manipulation that Frankfurt’s appeal to appropriate hierarchies of desires or attitudes is insufficient to account for morally responsible agency (Cuypers 2004). Whereas Frankfurt’s account of ultimate origination is internalist, my account is broadly externalist. Compatibilists respond differently to the manipulation problem and, accordingly, develop diverse theories about ultimate origination or agential participation.[7] On the plausible assumption that at least certain elements of a person’s psychology — especially, desires, preferences, values and other pro-attitudes — play an essential role in the ingredients for moral responsibility, *internalism* is the thesis that facts about how the person acquired these psychological elements in the past are completely irrelevant for his agential participation now. With certain qualifications, internalists claim that it does not, for example, matter whether the causal source of these elements is the result of manipulation, or of “natural” factors. *Externalism* is the thesis that facts about one’s past or history in the external world that bear on the acquisition of one’s psychological elements are pertinent to whether one’s actions are really one’s own and, hence, pertinent to whether one can be morally responsible for them. Again, with various caveats, externalists affirm that it does matter

[7] Internalism (or structuralism) is defended by, among others, Frankfurt 1988 and McKenna 2004. Externalism (or historicism) is defended by, among others, Christman 1991 and Mele 1995.

whether the causal source of the psychological elements is infected with manipulation, or is “natural” — and thus whether it is responsibility-subverting or possibly participation-preserving. This is not the place to adjudicate this in-house debate among compatibilists who divide into internalists and externalists.[8] For the purposes of this paper, I just note how my position differs from Frankfurt’s.

Whereas his internalist understanding of agent participation invokes decisive wholehearted identification, I understand agent investment in an externalist way as essentially associated with behaviour causally deriving from *authentic evaluative schemes*. I propose that the deep insight the Participation Principle, or principle O, captures may be expressed in terms of, what I call, the “Authenticity Principle”.

> *Authenticity Principle*: An agent is suitably “in touch” with an action of his — is properly “invested” in that action — only if the action causally stems from elements of an evaluative scheme of his that is authentic.

Condition (iii) of a compatibilist account of ultimate origination should, to my mind, be interpreted in terms of this principle. Only if the action causally stems from elements of an agent’s authentic evaluative scheme, would it be correct to say that the action “truly” issues *from the agent*, or is the “*agent’s own”,* or is one over which *the agent* has ultimate control. I now sketch the necessary background to substantiate this *Authenticity Principle*.[9]

[8] For my critique of internalism, see Cuypers 2004; for that of externalism, see Cuypers 2006.

[9] For the detailed picture, see Haji and Cuypers 2008, pp. 19-32)

4. Forward-looking relative authenticity account

4.1 The "hard" problem of origination

The problem of origination for externalist compatibilists is a problem about the actual causal histories of developing agents. David Zimmerman identifies the chief problem for causal history compatibilism as "... *the puzzle of naturalized self-creation in real time*: How do some children manage to develop the capacity to *make up their own minds* about what values to embrace, by virtue of having gone through a process in which they play an increasingly active role in *making their own minds*, a process that begins with their *having virtually no minds at all*?" (Zimmerman 2003, p. 638) How can an actual causal history, starting at birth, be such that it includes some non-alien-deterministic events so that the developing agent can properly be implicated, at times, in his actions?

It will be helpful to distinguish between two stages in an individual's development: the stage prior to which the individual has become a so-called "normative agent" — the *pre*-normative stage (before *t*) — and the *post*-normative stage (after *t*). We are not born as normative agents; we start off as non-normative beings and gradually develop into partially normative individuals until we finally become fully normative ones (at *t*). A (fully) normative agent is an individual capable of (1) intentional action, (2) rational deliberation or practical reasoning, and in the possession of (3) an evaluative scheme in the light of which he guides his deliberation and action. During the pre-normative stage an individual, as a child, gradually acquires an *initial* evaluative scheme. In the post-normative stage of childhood, adolescence and adulthood an individual maintains a scheme that results from modifications to his initial scheme — he possesses an *evolved* evaluative scheme. Such an evolved, or minimally completed initial scheme is made up of the following constituents: (i) normative standards the agent believes (though not necessarily consciously so) ought to be invoked in assessing reasons for action, or in evaluating beliefs about how the agent should go about making choices; (ii) the agent's long-term ends or goals he deems worthwhile or valuable — his "pro-attitudes"; (iii) deliberative principles the agent utilizes to arrive at choices about what to do or intentions how to act; (iv) and lastly, motivation both to act on the normative standards specified in (i) and to pursue one's goals of the sort described in (ii) at least partly on the basis of engaging the deliberative principles outlined in (iii). So, evaluative schemes contain both doxastic propositional attitudes (beliefs) and motivational elements (desires and other pro-attitudes).

The problem of origination is most pressing as regards the individual's acquisition of an initial evaluative scheme at the pre-normative stage. Most theorists, including myself, agree that changes in an already existing authentic evaluative scheme — an authentic evolved one — at the post-normative stage may be perfectly compatible with preserving authenticity on the condition that those changes take place under *the agent's own deliberative rational control*. Such self-control involves the exercising of deliberative capacities, including (a) the capacity critically to reflect on beliefs and pro-attitudes, (b) the capacity rationally and morally to assess these attitudes, and (c) the capacity to change their strength, or to revise and even to eradicate them, or to foster new attitudes in the light of (a) and (b). Exercising these deliberative capacities is authenticity-preserving in virtue of the agent's "engaging" elements constitutive of his authentic evolved scheme. By contrast, if an agent's authentic evaluative scheme is not engaged in, for instance, acquiring a pro-attitude — if its acquisition bypasses all of the agent's capacities of deliberative control — then this pro-attitude is inauthentic.

Whereas a solution along these lines of the problem of authentically modifying evolved schemes is relatively uncontroversial, theorists are in a quandary about the "hard" problem of origination — that of authentically acquiring initial schemes. How, precisely, does authenticity originate in the *initial* evaluative schemes of children who gradually develop into normative agents? Addressing the pre-normative agent stage — early childhood before a full-fledged evaluative scheme has been attained — is there a reasonable sense in which a child's cognitive and pro-attitudinal elements, constitutive of the initial scheme it will acquire, can be said to be authentic or "his own"?[10]

4.2 Authentic initial evaluative schemes

Regarding the child's initial evaluative scheme, I argue for the view that its constituent elements can be *relationally* authentic: they can be authentic relative to respecting or ensuring future moral responsibility. Elements of such a scheme can be, as I shall say, responsibility-relative authentic. Thus, my view on origination is *forward-looking*: although pertinent motivational elements instilled in the child during the educational process are not authentic *per se*, they can be authentic-with-an-eye-toward-future-moral-responsibility. Necessary interferences on the part of educators in the educational process are acceptable precisely

[10] There is a complication here: the acquisition of an evaluative scheme is a matter of degree; so, depending upon their stage of development, at various stages of maturation children may be partial normative agents. Does this mean that being the ultimate originator of one's actions or being morally responsible for them are gradual as well? My conjecture is that the conception of ultimate origination (authenticity) is discrete, while the concept of moral responsibility is gradual. An action either is or is not authentic according to the *Authenticity Principle*. Yet the authenticity condition is only one of several necessary conditions for moral responsibility. Since the fulfillment of these *other* conditions might be a matter of degree, moral responsibility can be gradual. Note that there are not only in childhood but also in adult life, at all times, always issues about being fully, or only partially morally responsible for certain actions.

insofar as they are required for the development of youngsters into morally responsible agents. Accordingly, the "hard" problem of origination is solved, first, by invoking the view that authenticity *per se* or *sans phrase* of an initial scheme's constituents is a myth and, second, by showing that things such as authoritarian indoctrination or harsh paternalism (when responsibility-thwarting), unlike authentic ways of instilling salient psychological elements, make use of ways that undermine such responsibility-relative authenticity.

To appreciate this strategy, reflect on the way in which mental illness, coercion, or deception affect moral responsibility. All parties readily grant that such factors frequently subvert moral responsibility. When they do so, they undermine one or more of the requirements of responsibility, such as epistemic or control requirements. If a person acts on the basis of a belief that is false (for example, the belief having been acquired as a result of deception), then (assuming that the person is non-culpably ignorant) the person is "off the hook". Similarly, if a person acts on a surreptitiously implanted desire that is irresistible, so that action issuing from the desire is action that is not under her control, then once again the person has a genuine excuse. Against the backdrop of these considerations, I propose that a cognitive or pro-attitudinal element or its mode of acquisition is *inauthentic* if that psychological element or the way in which it is acquired will *subvert* moral responsibility for behaviour, which owes its proximal causal genesis to the element, of the normative agent into whom the child develops. Subversion of moral responsibility would then occur as a result of either the epistemic or control requirement — *independently*, to avoid circularity, of the authenticity requirement itself — of moral responsibility being thwarted. Keeping in mind this proposal that instilment of doxastic and pro-attitudinal elements that subvert responsibility for subsequent, relevant behaviour suffices for the inauthenticity of these pro-attitudes, ponder these examples.

To be *morally* responsible for an action, an agent must be minimally morally competent. An agent must have elementary moral concepts, such as those of right, wrong and obligation, and he must be able to appraise morally — even if imperfectly — reasons, choices, actions, consequences of action (etc.) in light of the normative standards that are partly constitutive of his evaluative scheme. A minimally morally competent agent has a grasp of the notions of guilt, resentment, praise-, and blameworthiness or of the concepts of related reactive attitudes or feelings and has at least a rudimentary appreciation of when such attitudes or feelings are appropriate. Now suppose a child, call him "Émile"[11], is trained so that he lacks knowledge of the relevant moral concepts and norms with the result that he is not even minimally morally competent. Then lack of instilment of the appropriate moral concepts and norms is responsibility-subversive because without the conceptual wherewithal, Émile won't satisfy responsibility's epistemic requirement. Such required concepts and attitudes are then *authenticity demanding*. Or consider instilment in Émile of a pro-attitude or disposition — on a par with an irresistible desire — the influence of which on his behaviour he cannot thwart. Instilling such a pro-attitude would presumably undermine responsibility for later conduct arising from that pro-attitude by undermining the control moral responsibility requires. Or suppose instilled in Émile is a powerful disposition always to act impulsively. Here, again, we would not want to hold Émile morally responsible for much of his later impulsive behaviour. In sum, some interferences — untoward ones — are incompatible with Émile's being morally responsible for his subsequent behaviour which issues from these interferences. Such interferences subvert later moral responsibility while others do not. I propose that the subversive ones are responsibility-relative inauthentic and that the ensuing attitudes are thus *authenticity destructive*. Setting aside the authenticity requirement of responsibility, if

[11] After Jean-Jacques Rousseau's hero in *Émile ou de l'éducation* (1762).

these interferences subvert later moral responsibility, they will do so by subverting *other* requirements of responsibility, such as epistemic or freedom requirements.

So far, our discussion has been limited to responsibility-relative authenticity of the "objects" of instilment such as dispositions, or pro-attitudes in general. What about the methods or techniques of instilling such things; are some responsibility-relative authentic and others not? Assume that to ensure prevention of subverting moral responsibility for later behaviour, it is necessary to instil in the child the disposition to be moral. Different modes of instilment of this disposition could affect responsibility-relative authenticity of this very disposition itself. For example, suppose that given the mode of instilling the moral disposition in Émile — perhaps the disposition was "beaten into" Émile, or instilled via "shock therapy" — Émile subsequently finds that he cannot refrain from doing what he perceives to be morally right and to do what is, for instance, in his best self-interest. On occasions of choice, he is stricken with inward terror even at the faintest thought of not doing what he deems moral. Intuitively, Émile would not be morally responsible for much of his later behaviour because the mode of instilment of the moral disposition subverts responsibility-grounding control. Modes of instilling pro-attitudes (etc.) are responsibility-relative not "truly one's own" (that is, are responsibility-relative inauthentic) if the modes subvert responsibility for later behaviour. Again, if these modes of acquiring pro-attitudes undermine later moral responsibility, they will do so by subverting one or more of responsibility's requirements (other than the authenticity requirement). These modes of attitude-acquisition are then *authenticity subversive*.

Apart from pro-attitudes, a normative agent's evaluative scheme comprises *cognitive* constituents. Again, with the young child whose evaluative scheme is in embryo, it may well be the case that certain beliefs will have to be wilfully instilled to ensure responsibility-relative authenticity. One will, perhaps, have to instil in Émile, for example, the belief that critical self-evaluation is important because without this belief, moral responsibility for his later behaviour may well be threatened in the manner indicated above. In addition, Émile's having of such a belief, it would seem, would be morally permissible and perhaps even morally required. Instilling beliefs of this sort, in consequence, via modes or methods that themselves do not subvert later responsibility, would not threaten responsibility-relative authenticity. Various sorts of belief, though, would undermine or seriously imperil moral responsibility for later conduct. The following sorts, for example, seem to be responsibility-relative inauthentic: beliefs formed as a result of deception (and self-deception), beliefs implanted on the basis of coercive persuasion or subliminal influencing, and beliefs inculcated in such a way that the agent is subsequently never encouraged to seek supporting evidence for them and his reason assessment capacity is permanently suppressed. Émile, presumably, would not be morally responsible for actions performed in the light of such beliefs.

4.3 A compatibilist authenticity criterion

I propose, then, the following criterion as one that governs responsibility-relative authenticity of *initial* schemes of developing agents at the pre-normative stage.

> *Authenticity Criterion*: A child's initial evaluative scheme is responsibility-relative authentic if its doxastic and pro-attitudinal elements (i) include all those, if any, that are required to ensure that the agent will be morally responsible for its future behaviour; (ii) do not include any that will subvert the agent's being responsible for future behaviour that issues from these elements; and (iii) have been acquired by means that, again, will not subvert the agent's being responsible for its future behaviour.

All the ingredients for giving a solution to the "hard" problem of origination are now in place. To ensure that the child matures into a normative agent, certain doxastic and pro-attitudinal elements must be instilled in the child. Instilling pertinent beliefs or desires is authentic if their acquisition does not subvert, in a characteristic way, moral responsibility for later behaviour that (at least partly) issues from these elements. The characteristic way is this: The acquisition of these elements subverts moral responsibility by compromising necessary requirements of responsibility, such as epistemic or control ones, with the exception of the authenticity requirement itself. These elements are, then, in the terminology introduced, relative-to-future-responsibility authentic and gradually build up a child's authentic initial evaluative scheme in compliance with the *Authenticity Criterion*. But some instilled elements or their modes of instilment undercut moral responsibility for later behaviour by undermining fulfilment of necessary conditions of responsibility other than the authenticity condition itself. Offensive manipulation, harsh paternalism, hideously depraving conditions, or experiences traumatic to the child may have this effect. If they do (and empirical evidence is required to confirm whether they do), then in these sorts of case, the instilled elements are (relationally) inauthentic — not "truly the child's own".

Hence, my earlier *Authenticity Principle* of compatibilist origination must be interpreted in terms of this *Authenticity Criterion*. An agent is properly "invested" in an action of his — is the ultimate originator of that action — only if the action causally springs from elements of an evolved evaluative scheme that is based on an initial evaluative scheme, both of which schemes are authentic in accord with the pertinent criteria. An evolved evaluative scheme is authentic when its changes in the past took place under the agent's own deliberative control, while an initial evaluative scheme is authentic when it complies with the *Authenticity Criterion*.

5. A Compatibilist response to Pereboom

In response to Pereboom, and especially the second premise (2.) of his argument for hard incompatibilism, I submit that the *Authenticity Criterion* can make a relevant and principled distinction between alien-deterministic events and authentic-deterministic ones. According to Pereboom, all covertly manipulated action-producing events, such as decisions, are alien and, because there is no significant difference between ordinary causation and manipulation, all causally determined actional elements are alien too. This is, however, tendentious. Not all causation, and not even all manipulation, is menacing in the sense that it is responsibility-undermining. Let me explain.

Assume that all changes in an evolved evaluative scheme took place under the agent's own deliberative rational control, so that all of them were authenticity-preserving. Action-producing events causally stemming from elements of this evolved scheme are then authentic and not alien, if the causal history of these elements traces back to constituents of his *authentic* initial evaluative scheme. Now whether these original constituents are ordinarily caused or manipulatively caused does not matter *in se* — authenticity *per se* is a myth. The only thing that counts, in accord with the *Authenticity Criterion*, is whether or not they undermine responsibility for later behaviour by undermining epistemic or control requirements of responsibility. Hence, if the original constituents were ordinarily caused but nonetheless undermined these requirements, then they would be alien despite the fact of their "natural" origination. Alternatively, if these elements were manipulated *without* undermining the epistemic or control requirements, then they would be authentic after all and we would have a case of innocuous manipulation. So, using the term "manipulation" with no distinction between normal (or "baseline") and deviant causal chains in mind, as Pereboom does, is question-begging against the compatibilist.

In sum, by appealing to the *Authenticity Criterion*, the compatibilist can very well draw a principled distinction between authentic-deterministic and alien-deterministic events, or for that matter between innocuous and menacing "manipulation". Consequently, in a deterministic world all events are indeed deterministically caused, but some are alien-caused and some are authentic-caused in accordance with the *Authenticity Criterion*.[12]

Katholieke Universiteit Leuven
Institute of Philosophy
Email: Stefaan.Cuypers@hiw.kuleuven.be

REFERENCES

Berofsky, B., 2006, "The Myth of Source", *Acta Analytica*, **21,** 3-18.

Christman, J., 1991, "Autonomy and Personal History", *Canadian Journal of Philosophy* 21, 1-24.

Cuypers, S. E., 2004, "The Trouble with Harry: Compatibilist Free Will Internalism and Manipulation", *Journal of Philosophical Research*, **29**, 235-54.

Cuypers, S. E., 2006, "The Trouble with Externalist Compatibilist Autonomy", *Philosophical Studies*, **129**, 171-96.

Fischer, J. M. and Ravizza, M., 1998, *Responsibility and Control: An Essay on Moral Responsibility*, Cambridge University Press, Cambridge.

[12] For our application of the *Authenticity Criterion* to Pereboom's four-case argument, see Haji and Cuypers 2008, pp. 40-1.

Frankfurt, H. G., 1988, *The Importance of What We Care About*, Cambridge University Press, Cambridge.

Haji, I., 1998, *Moral Appraisability. Puzzles, Proposals, and Perplexities*, Oxford University Press, Oxford.

Haji, I., 2009, *Incompatibilism's Allure. Principal Arguments for Incompatibilism,* Broadview Press, Peterborough.

Haji, I. and Cuypers, S. E., 2008, *Moral Responsibility, Authenticity, and Education*, Routledge, New York.

Kane, R., 2002, "Introduction: The Contours of Contemporary Free Will Debates", In *The Oxford Handbook of Free Will*, 3-41, R. Kane, (ed.), Oxford University Press, New York.

McKenna, M., 2004, "Responsibility and Globally Manipulated Agents", *Philosophical Topics*, **32**, 169–92.

Mele, A. R., 1995, *Autonomous Agents. From Self-Control to Autonomy*, Oxford University Press, New York.

Pereboom, D., 1995, "Determinism *al Dente*", In *Free Will*, 1997, 242-72, D. Pereboom, (ed.), Hackett Publishing Company, Indianapolis.

Pereboom, D., 2001, *Living Without Free Will*, Cambridge University Press, Cambridge.

Pereboom, D., 2002, "Living Without Free Will: The Case For Hard Incompatibilism", In *The Oxford Handbook of Free Will,* R. Kane, (ed.), 477-88, Oxford University Press, New York.

Pereboom, D., 2004, "Is Our Conception of Agent-Causation Coherent?", *Philosophical Topics*, **32**, 275-86.

Pereboom, Derk. 2007. "Hard Incompatibilism", In *Four Views on Free Will,* J.M. Fischer, R. Kane, D. Pereboom, and M. Vargas, 85-125, Blackwell, Oxford.

Smith, H., 1983, "Culpable Ignorance", *The Philosophical Review,* **92**, 543-571.

van Inwagen, P., 1983, *An Essay on Free Will*, Clarendon Press, Oxford.

Velleman, J. D., 1992, "What Happens When Someone Acts?", *Mind,* **101**, 461-81.

Widerker, D. and McKenna, M. (Eds.), 2003, *Moral Responsibility and Alternative Possibilities*. *Essays on the Importance of Alternative Possibilities*, Ashgate, Aldershot.

Zimmerman, D., 2003, "That Was Then, This Is Now: Personal History vs. Psychological Structure in Compatibilist Theories of Autonomous Agency", *Noûs*, **37**, 638-71.