Philosophica 85 (2012) pp. 67-89

MORALLY EMBEDDED SELVES AND EMBEDDED COMPATIBILISM

Guy Pinku[*]

ABSTRACT

The principal argument suggested here is that we are all morally embedded selves: We have no control over the abilities that make us moral agents nor can we control the degree to which we have these abilities; in other words, we are not responsible for our good or bad qualities as moral agents. This, I believe, calls for the adoption of embedded compatibilism (EC). According to EC, people have

[*] I would like to thank Saul Smilansky for long periods of generous guidance; Shaun Gallagher for hosting me in UCF (where I was able to finish writing this article); Josef Horowitz, Arnon Cahen, David Kaplan, Roy Sorenson and Shaun Gallagher for reviewing various drafts of this paper; Mason Cash for his comments; Marion Lupu for linguistic editing; Tom Claes, Tim De Mey, Marijke de Pous and two anonymous reviewers for final revisions; Maureen Sie, Arno Wouters and Nicole van Voorst Vader-Bours for inviting me to give a lecture based on a draft of this paper; Annemarie Kallis for her helpful commentary on this lecture. I would also like to thank the participants of the workshop on Strawsonian and consequentialist views on personal responsibility which took place the 15th and 16th of October 2010 at Erasmus University, Rotterdam (where the debate between Derek Pereboom and Manuel Vargas concerning basic desert was an inspiration for me).

control over their conduct; this control, however, is embedded within prerequisites, which they cannot control and hence are not responsible for having or lacking. On the one hand, EC enables us to explain why a lack of control at the ultimate level does not eliminate moral judgment altogether. However on the other hand, EC ought to change our understanding of moral responsibility; inter alia, it supports a hybrid notion of punishment, indicates the incomplete nature of guilt and reintroduces the problem of moral luck.

1. Morally embedded selves

I should like to begin by presenting some background material. The compatibilist accepts the lack of libertarian free will and therefore looks for a criterion—that is, a characterization of the agent—which can justify conferring moral responsibility on to her or him. The criterion that the mainstream compatibilist endorses is reason; her approach is, thus, ‘reason-based compatibilism.’ An influential version of reason-based compatibilism is Fischer and Ravizza's (1998) notion of reason responsiveness. Fischer and Ravizza suggest that moral responsibility is based upon reason responsiveness: the ability to recognize reasons for a behavior and to act upon them.

A reason responsiveness mechanism, however, assumes various psychological capacities and psychological inclinations which may include, inter alia, an ability to understand other persons' mental states, a tendency to self-reflection and self-analysis, intelligence, and abilities of self-control.[1] Also, various external circumstances such as emotional stress or even fatigue may influence one's reason responsiveness abilities. The main point suggested here is that people cannot control all or even most of these; hence, people *cannot* control the abilities that make them moral agents nor can they control the degree to which they possess these

[1] I also assume here points suggested by Richard Double (1991, Chap. 2).

abilities; briefly, a person is not responsible for her (or his) good or bad qualities as a moral agent. In other words, a reason responsiveness mechanism does not settle the issue of moral responsibility completely, some residual difficulty still remains.

A critic may suggest, however, that a person can determine the content of the moral norms which she or he acts upon. I do not agree with this statement since it suggests that a person has more control than she or he actually has. I may begin with the observation that norms of morality, to some degree at least, are relative to the society in which they are present. A child in the process of becoming a moral agent learns the 'adequate' norms of morality, namely, the socially accepted norms. An example could be the group of agents whose interests one should take into consideration, which interests ought to be taken into consideration, and when. For instance, suppose that in the distant future the norms concerning the eating of meat will be different from those of today and that meat eating will be considered as wrong almost as one regards cannibalism nowadays. This, however, does not make people who eat meat *nowadays* morally corrupt. Most of humanity is educated according to moral norms that do not condemn meat eating, hence people have a limited ability to recognize the wrongness of their actions in this realm - if indeed meat eating is morally corrupt. A similar point suggested by Russell (2002) is that one cannot determine the kind of reason responsiveness mechanism that one owns.

However, it might be suggested that intentional self-reflection (such as that which adolescents are prone to have, or such as that which people may have at the beginning of a new year) may enable a person to choose the values and the moral norms that she or he lives with. I do not underestimate the significance of self- reflection; however, human beings are usually not involved with self-reflection as a daily habit; that is to say, the default mode of the human condition is acting rather than reflecting.[2] Moreover, even if indeed people were prone to be more self-reflective, self-reflection cannot reach the degree of self-creation. Unavoidably some inclinations, values and ways of thinking ought to be uncritically assumed within the process of self-reflection. Moreover, in every culture there is probably a danger of unnoticed moral blind spots.[3] In the 1850s, for example, Afro-Americans were not considered as persons whose interests should be taken into consideration. Hence, the slave-owners in the 1850s, cannot be blamed as responsible for deprivation of human rights, since it is doubtful whether they were (and could be) fully aware of the wrongness of their actions (Wolf, 1987/2003). The point is that a person, at least to some extent, cannot control the moral norms upon which she or he acts and hence personal blame in cases clearly assosiated with corrupted moral norms is elusive.[4]

[2] I have in mind the Monty Python script of a football match between philosophers; a match in which the players, instead of kicking the ball, walked in the football pitch going through a series of endless questions.

[3] I follow here Thomas Nagel (1986): "But even allowing for unlimited time, or an unlimited number of generations, to take as many successive steps as we like, the process of enlarging objectivity can never be completed, short of omniscience. First, every objective view will contain a blind spot, and cannot comprehend everything about the viewer himself. But second, there will not even be a limiting point beyond which it is impossible to go. This is because each step to a new objective vantage point, while it brings more of the self under observation, also adds to the dimensions of the observer something further which is not itself immediately observed. And this becomes possible material for observation and assessment from a still later objective standpoint. The mind's work is never done." (pp. 128-9)

[4] Wolf's example (1987/2003) may lead to controversy. Marijke de Pous (private correspondence) has pointed out that there is a gap between arguing that slave-owners were not in control of the moral norms they grew up with and concluding that they cannot be blamed as accountable. I believe, however, that it is doubtful

There is another aspect to the relativity of the moral norms. There might be circumstances in which an agent, who has adequate psychological abilities, *nonetheless* holds moral norms that deviate from the socially accepted ones; there are, for example, being brought up by non-normative caregivers such as delinquent parents or cases of immigrants having grown up in different subcultures (Wolf, 1990). Hence, '*normative adequacy*', the extent to which the agent's norms of morality are in harmony with the norms accepted by society, constitutes an *additional* factor that may determine the person's moral abilities (due to determining their adequacy); normative adequacy is also not controlled by the person.

Therefore we may conclude that we are all morally embedded selves: We have *no control* over the abilities that make us moral agents, and our ability to control our normative moral profile is quite limited.

1.1 Distributed cognition

The notion of morally embedded selves is in line with the notion of distributed cognition which is suggested within the realm of cognitive science.

that a person can grasp the wrongness of actions associated with durable corrupted moral norms; *for this reason*, she or he cannot be blamed for such actions (see also part 1.2). Yet, there are two further qualifications: (1) There is a distinction between taking responsibility and responsibility which is associated with blame and punishment, e.g., one may argue that a person ought to take responsibility even when she or he cannot be blamed as accountable. (2) Personal moral responsibility and culpability are only part of the considerations that may justify punishment (see part 3.1 and Levy, 2003). I wish to thank Marijke de Pous and an anonymous reviewer for their comments on this issue.

According to Pettit (2007), for example, control is displayed by a capacity of ‘conversability;’ namely, an ability to discuss “a currency of reasons for thought and action that are recognized as relevant on all sides [of a conversation]” (p. 82). This sort of responsiveness to reasons implies a *holistic ability* of control; that is, an ability to explain one's actions (by reasons) and to modify them according to reasons rather than an ability to initiate directly each act that one takes. According to this analysis the agent is an editor rather than an author, or a conductor rather than a composer; hence, Pettit (2007) suggests: “Although unthinking habit shapes what agents do, the discipline of reason will be in virtual control so far as ready to be activated and take charge in the event of habit failing to keep the agent in line.” (p. 84).

Andy Clark (2007) portrays a similar picture of the agent, based upon the notion of ecological control. Ecological control is a strategy of control that avoids micromanaging by taking for granted various factors of the environment and characterizations of the organism's body. One example of this would be a computation of walking that takes for granted various environmental conditions, such as the shape of the walker's body, gravity force, the steepness of the path and, by virtue of this, reduces the amount of computation needed to control walking.

According to this picture, *self*-control should *not* be identified with a conscious deliberate reasoning, but rather with all the processes, conscious and unconscious, that constitute control. With regard to this, Clark (2007) cites Dennett's metaphor:

> “Our free will, like all our other mental powers, has to be smeared out over time, not measured at instances. Once you distribute the work done … in both space and time in the brain, you have to distribute moral agency around as well. You are not out of the loop; you are the loop.” (Dennett, 2003, p. 242; cited by Clark, 2006, p.108)

This analysis suggests that deliberate conscious reasoning is only one aspect of the self but not '*the* self.' Clark's additional step is to distribute control outside of the brain, and include within the controlling system (or even within the self)[5] environmental factors such as 'extra-neural stores, strategies, and processes.'

1.2 Distributed cognition and morally embedded selves

Morally embedded selves and distibuted cognition both imply a weak notion of self-control and assume externalism.

A weak notion of self-control: Distributed cognition implies reduction of the centrality of the conscious reasoning aspect of the self due to a distribution of cognitive processes between various mechanisms whereas the notion of morally embedded selves suggests that the causal chain continues beyond the 'conscious reasoning self' (e.g., the self cannot control the qualities that make her or him a moral agent). Hence, distributed cognition and morally embedded selves both imply the falsity of the image of a conscious reasoning self as one single ultimate cause of behavior. They suggest, instead, that there is a conscious reasoning aspect of the self which is capable of explaining behavior and modifying some of it.

Externalism: The person's social environment ought to be understood according to the notion of ecological control; in other words, the person's social environment is one of the mechanisms (such as extra-neural stores,

[5] Clark (2007) suggests a notion of soft selves: "We are "soft-selves," continuously open to change and driven to leak through the confines of skin and skull, annexing more and more nonbiological elements as aspects of the machinery of mind itself." (p. 112)

strategies, and processes) that, according to the notion of ecological control, take part in the control loop of the agent's behavior.

A child, within the socialization process, absorbs norms (including moral norms) that accommodate her or his interpersonal interactions such as acting according to a system of socially-accepted commitments. It is quite clear that a child accepts these norms relatively uncritically. A mature person may examine or even re-establish some of these norms; however, this process, as we said earlier, is quite limited. *Hence*, it is often the case that the justification and the causes for the individual's moral norms are ‘external,’ they might be found, to give some examples, within the social level of the accepted moral systems, social factors, historical circumstances and even biological influences. This implies that the social environment takes part within the non-biological factors that constitute ecological control. Or, in other words, within a delineation of an individual's *self*-control the social environment should be included. There is no point to moral judgment of the (individual) agent unless the moral norms with which she or he was raised are assumed; hence, appreciation of praise and blame ought to assume the ‘external’ normative background.

2. Embedded compatibilism (EC)

The libertarian view consists of two claims: (1) Moral responsibility is *incompatible* with a lack of libertarian free will; however (2) We do have libertarian free will. Rejection of the second claim alone leads to hard determinism whereas the rejection of both leads to compatibilism; in other words, compatibilism is the view that we lack libertarian free will and that this is compatible with having moral responsibility.[6]

[6] Compatibilism may also originate from assuming indeterminism (at least from most sorts of indeterminism; Pereboom, 2001). Hence, it is better to characterize

The notion of morally embedded selves leads to a rejection of the libertarian view (since the notion of libertarian free will suggests that the individual person serves as an ultimate cause for her or his behavior). However, it seems that the notion of morally embedded selves is in harmony both with compatibilism and with hard determinism. I suggest, however, that it justifies what I call *embedded compatibilism* (EC). Presenting an argument for this proposal may also serve as an initial sketch of EC.

Henceforth I shall defend the notion of EC by ‘a voyage’ through three intermediate stages. I may begin with what might be called *pessimistic hard determinism*, move to Smilansky's (2000) *fundamental dualism*, and then to Pereboom's (2001) *(half) optimistic hard determinism.*

According to the pessimistic hard determinist, the lack of libertarian free will implies that people's behavior consists only of unfolding the cards they already had, hence people *do not* deserve any credit for their behavior. In other words, nothing in the action of a person ‘belongs’ to the person herself *in virtue of self-control*, since what is called the person's behavior is all determined, actually, by initial conditions which the person cannot control. According to this analysis, the responsibility system (i.e., notions such as desert, practices such as reward and punishment, and reactive attitudes such as resentment and gratitude) loses its grip. The notion of morally embedded selves, however, suggests that this analysis goes too far; *within* a context of ‘given’ abilities and social norms there is no reason to assume that people lack control of their conduct.[7] In other words, people do have control of (and responsibility for) their actions, though they cannot control possession of the capability

compatibilism as the view that moral responsibility is compatible with a lack of libertarian free will.

[7] Later, Pereboom's (2001) argument of basic desert is rejected.

for moral agency and they have only a weak control over its type and its quality.

This analysis may lead to Smilansky's *fundamental dualism* (2000). Smilansky suggests that at the ultimate level there is neither control nor responsibility; in addition there is, however, a "[…] level of local compatibilist freedom or control, without enquiry into the ultimate level. [This is] the level of analysis at which it might be correct to say that agents have free will in a way relevant for moral responsibility, even though they do not have libertarian free will." (p. 313).

Hence, Smilansky argues that (often) there is a tension between two different levels of analysis: the local level suggests that the agent *does* deserve praise or punishment whereas the ultimate level suggests that she *does not*. Moreover, Smilansky suspects that revealing this tension would undermine the compatibilist level. He proposes, therefore, to avoid revealing the libertarian illusion which he assumes that people have with regard to the ultimate level. A metaphor might be helpful here: One may imagine the belief in libertarian free will as a beast that carries the compatibilist level - the level at which the personal interaction takes place, so Smilansky calls to preserve the libertarian beast in order to protect its compatibilist rider.[8]

Pereboom (2001) is a (half) *optimistic hard determinist*. Pereboom suggests that respect for persons and reactive attitudes in most of their aspects are not undermined by hard determinism; that is to say that hard determinism is in harmony with respect for persons due to their rational capacities (and that the same applies to most of reactive attitudes).[9] Thus,

[8] This metaphor was suggested by Smilansky (private correspondence).

[9] For example: "Achievement and life-hopes are not obviously tied to praiseworthiness in the strong way […]. If one hopes for a certain outcome, then if one succeeds in acquiring what one hoped for, intuitively this outcome can be one's achievement, albeit in a diminished sense, even if one is not praiseworthy for it." (Pereboom, 2001, p, 194)

Pereboom's approach implies that although the capacities of a person are not controlled by her, there is still a reason to respect her because she *has* these capacities. So, it might be suggested that the compatibilist view that there is an important distinction between a deliberate action and an unintended action (due to a spasm, for example) is also assumed by Pereboom *in regard to respect for persons and reactive attitudes* (Unless such a distinction is assumed, there is no point in respecting people for their rational capacities).

However, Pereboom (2001) makes a sharp distinction between respect for people due to their rational capacities and moral responsibility. This distinction is grounded upon the assumption of basic desert:

> "[…] in my view, for an agent to be *morally responsible for an action* is for this action to belong to the agent in such a way that she would deserve blame if the action were morally wrong, and she would deserve credit or perhaps praise if it were morally exemplary. The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve the blame or credit just by virtue of having performed the action, and not, for example, by way of consequentialist considerations." (p. xx)

There might be two versions of basic desert: (1) *The strong version:* Desert is associated only with the agent; any external consideration is irrelevant. (2) *The moderate version:* Desert is associated only with the agent; however the notion of agent itself assumes a context which the

"The hard incompatibilist need not deny that human beings are rational and responsive to reasons, and no feature of her view threatens the respect she has for them because of their rational capacities." (Pereboom, 2001, p, 206) agent cannot control. I believe that if Pereboom had adopted the moderate notion of basic desert then he would not have assumed a sharp distinction between respect for people due to their rational capacities and desert for people because of their deliberate actions (i.e., the moderate version does not suggest that the lack of control of the prerequisites needed for desert makes desert inapplicable).

The distinction between the moderate and the strong versions of basic desert suggested above assumes a distinction between embedded desert and ultimate desert. Embedded desert is applicable only within certain conditions, whereas ultimate desert relates to what a person deserves without assuming any context at all. *EC assumes embedded desert.*

A critic may ask: What is the difference, then, between the suggested notion of EC and Smilansky's fundamental dualism? The EC may reply that Smilansky (2000) does not make a distinction between ultimate and embedded desert; therefore he suggests that (often) moral judgments face unbearable tension between levels of analysis. But, the notion of embedded desert enables one to reduce this tension; within embedded desert there is no inherent tension.

Smilansky may suggest, however, that there is still a tension between ultimate desert and embedded desert. The EC may reply that since we are morally embedded selves the notion of ultimate desert does not apply – it does not make sense to judge an agent without assuming the pertinent psychological abilities and social normative background, so we stay only with embedded desert. However, Smilansky may suggest that this reply consists of changing the subject: the EC relates to a different notion of desert and in virtue of this assumes a lack of tension. The EC may agree that the subject of the discussion has been changed: embedded desert is indeed a different notion from ultimate desert (I shall discuss the differences henceforth by relating to an additional criticism).

There might be a criticism from the opposite side. It might be argued that there is no difference between EC and compatibilism in general. I suggest putting the finger on the difference between EC and compatibilism by considering Fischer's (2007) reply to the argument that the ultimate level of analysis undermines moral responsibility:

> “Imagine here that our agency is represented by a horizontal line-segment from point b to point c. This is the Agency Line. Now suppose there is a vertical line coming from below, with an arrow pointing toward the Agency Line. The vertical line represents a causally necessary condition (or enabling condition), such as the sun’s shining; the sun’s shining causally sustains and “sets the stage” for the existence of agency. Now add a line that is (like the Agency Line) horizontal, starting to the left of point b at some point a, connecting a and b, and with an arrow pointing towards b. Suppose that the relevant agent is not in control of this antecedent causal sequence “pointing horizontally toward b,” just as he is not in control of the sun's continuing to shine. My question is: what is the difference between the vertical and horizontal lines? More carefully, if one is not troubled by the existence of the vertical line, why be troubled by the horizontal line? The two lines are equally “external” to the Agency Line, and thus mere appeal to internality will not distinguish the lines.” (p. 69)

Fischer suggests that the ultimate level makes no difference. The (pessimistic) hard determinist suggests that only the ultimate level ought to be considered and hence the responsibility system is inapplicable whereas Smilansky’s fundamental dualism suggests that both the ultimate level and the local level ought to be considered. My suggestion, based upon a revisionist strategy proposed by Manuel Vargas (2004; 2007), is to revise the vague common-sense notion of desert and to accommodate it to our actual state as morally embedded selves. This ‘new conceptual tool,’ I believe, may enable us to avoid the contrastive analyses at which

Smilansky points. This implies that the notion of EC ought to accept the incompatibilist insights concerning lack of control and responsibility at the ultimate level.

A critic may suggest, however, that following the revisionist strategy is an arbitrary choice. The EC may reply that also adhering to the current state (of some sort) of fundamental dualism is a question of choice. In any event, we have to choose. The question is: what is our justification? As suggested earlier, EC is justified due to its correspondence to our actual state as morally embedded selves. Moreover, the disadvantages of the alternative may support EC. As we shall see very soon, the main disadvantage of the current state (i.e., holding compatibilist practices by retaining the illusion of libertarian free will) is that it masks distortions in our moral judgments. So, if we have a viable alternative, one that supports our compatibilist practices while reducing some related distortions, we have to adopt it.

3. EC and punishment

The belief that we *have* libertarian free will is part and parcel of the common-sense picture we have of ourselves in a variety of circumstances. When we are making decisions, for instance, we tend to believe that the full range of possibilities is open, i.e., that we can turn in any direction and that there is nothing to hinder us from deciding freely (See also Smilansky, 2000 and Vargas, 2007). This analysis suggests that internalization of the lack of libertarian free will should change, at least somewhat, our common-sense notion of a self (or a person). It could, for instance, undermine the notion of respect for persons insofar as it involves an assumption of *ultimate* desert; it could also undermine the natural tendency for moral self-complacency, or reduce or even eliminate the tendency of inflationary feelings of admiration or hatred toward other persons, which is associated (in my opinion) with a false belief in the plausibility of moral saints or moral monsters; i.e., persons who are the ultimate sources of their bad or good will. These and other ideas should take part within a fully fledged delineation of EC. Here, however, I will illustrate the notion of EC by its analysis of punishment.

According to EC, people *have* control of over their conduct; this control, however, is embedded within prerequisites of which they cannot control and hence are not responsible for having or lacking. For this reason EC *cannot* support any sort of one-sided justification of punishment; namely, justification of punishment only according to desert, i.e., retribution or, on the other hand, justification of punishment without assuming desert at all, for example, pure utilitarian justification or justification by an analogy between punishment and quarantine (Pereboom, 2001).

EC takes the notion of retribution with a pinch of salt. According to EC, a lack of libertarian free will suggests that we *cannot* retain "moral responsibility of such a kind that, if we have it, it makes sense, at least, to suppose that it could be just to punish some of us with (eternal) torment in hell and reward others with (eternal) bliss in heaven." (G. Strawson, 2003, p. 216; cited by Bomann-Larsen, 2010, p. 2). According to EC, our concept of punishment ought to be limited and partial. A person might indeed deserve punishment; however, there is no point in thinking about guilt or praise in ultimate terms; any moral judgment is limited to a certain *assumed* scope of social and psychological circumstances.

The characterization of deserving punishment as a relational feature is *in line* with situating the general justification for the institution of punishment within the social level. Hart (1970), for example, distinguishes between the general aim of punishment (i.e., to set up standards of behavior) and the question of to whom a punishment may apply (i.e., restriction of punishment to morally responsible offenders).[10]

[10] Vargas (2004; 2007), complementarily, articulates a general principle: beliefs, emotional attitudes and practices pertinent to moral responsibility (i.e., the moral

The distinction between two levels of justification for punishment (i.e., the social level and the level of the individual agent) reflects *the complexity of guilt*: It is true that most of the people are moral agents who deserve to be punished when they offend moral standards and/or the law; yet, this applies only certain *assumed* or ‘given’ prerequisites of social norms and psychological abilities. In other words, desert is not applicable beyond some assumed framework (one may compare this to imagined possible worlds in which different laws of physics might be applied); however, because we are so immersed within our own framework we tend to have a (false) impression that features that are limited to it are ultimate, and hence that desert may justify ultimate guilt or, metaphorically, that a person might be guilty ‘in front of God.’ Complementarily, the notion of the offender as a moral monster, i.e., as if she were the *ultimate* creator of her own mean character and intentions is inapplicable, since this implies that no framework of circumstances could be assumed; in this regard, people are not ‘ultimately’ bad or good.

I may imagine a critic arguing that the EC analysis of punishment is perplexing. In order to convict an offender we need a definite position: *Either* a person deserves to be punished and hence might be guilty or a person does not deserve to be punished and hence cannot be guilty. The EC reply is that she does not suggest that a person is guilty and not guilty simultaneously, but rather that she holds a revised notion of desert and guilt.

The critic may agree that the ultimate levels of guilt and blame do not apply, but might still maintain that this does not make a diffrence, since embedded desert *actually* functions as ‘ultimate desert;’ so the new suggested titles ‘embedded desert’ and ‘partial guilt’ are merely titles. They do not solve the tension that the fundamental dualism presents. The

system) are justified by their effect: Fostering agents to mold their conduct according to moral reasons.

EC's reply is that embedded desert is not merely a 'new title' but rather it is a revised *notion* of desert; a notion that is different from the common-sense notion of desert. Hence, it leads to a different understanding of moral responsibility: inter alia, it enables one to reduce the tension that the fundamental dualism presents. Another result that embedded desert leads to is indicating the incomplete nature of guilt, which may lead to a notion of punishment that *does not* assume ultimate guilt of nor an associated emotion of hatred for 'the moral monster,' hence, internalization of the notion of embedded desert may lead to a degree of emotional equanimity concerning blame and guilt and may undermine the motivation for an extreme degree of punishment such as the death penalty. In addition, as we shall soon see, the notion of embedded desert leads to an acknowledgement of the problem of moral luck *within the framework of compatibilism*.

A further criticism may relate to practicality. It might be suggested that even if we assume that the notion of embedded desert is coherent, it may lead to confusion, and therefore we ought to retain the notion of ultimate desert (remember the beast and the rider metaphor). I believe, however, that people *are able* to internalize explicitly the delicate notion of embedded desert. I should like to suggest two examples.

Herbert Morris (1968)[11] suggests that therapeutic procedures which circumvent reason rather than address it are problematic since they do not respect the autonomy of persons, e.g., alleviating a tendency for bouts of violent and explosive anger by taking a drug such as Prozac. Pereboom (2001) replies to this criticism by arguing that "this sort of treatment often produces responsiveness to reasons where it was previously absent. A person beset by violent and explosive anger will typically not be responsive to certain kinds of reasons, to which he would be responsive if he were not suffering from this problem." (p. 180). I believe that the dialectic here highlights the point that people are morally responsible and

[11] Here, I follow a discussion of Pereboom (2001), pp. 179-180.

autonomous only when certain prerequisites are fulfilled. This analysis suggests that respect for persons is indeed precious; however, it is coupled with an awareness of some limitations. Morris's criticism of *any* circumventing therapeutic procedures is not convincing since it ignores the prerequisites needed for regarding people as moral agents and as autonomous.

A second instance relates to the realm of the interpersonal. This example originates from Strawson's (1962/1974) distinction between having emotional attitudes and having objective attitudes toward people.[12] We may get the impression that these attitudes are mutually exclusive; e.g, treating an insane person from an objective point of view is bound up with a suspension of one's emotions toward that person and adopting a non-judgmental 'emotionally cold' understanding. However, there are cases in which we move swiftly from one attitude to another with *regard to the same person*.

> "We look with an objective eye on the compulsive behavior of the neurotic or the tiresome behavior of the very young child, thinking in terms of treatment or training. But we can sometimes look with something like the same eye on the behavior of the normal and the mature. We have this resource and can sometimes use it: as a refuge, say, from the strains of involvement; or as an aid to policy; or simply out of intellectual curiosity." (pp. 9-10)

I suggest that this movement between the attitudes is typical of our interpersonal conduct; that is to say, in many cases we move swiftly, partly unconsciously, from emotional attitudes to objective attitudes and vice versa. There are, for instance, cases in which we realize that the

[12] An objective attitude may lead to treat persons "…as a subject for…treatment; … to be managed or handled or cured or trained;" (Strawson, 1962/1974, p. 9).

other person is in a weak state of self control (e.g., due, possibly, to exhaustion or anger); or there are cases of insolvable controversy in which (‘as a refuge, say, from the strains of involvement,’ as Strawson puts it elegantly) we may choose to attribute the person's approach to the circumstances of her or his life. This observation, I believe, suggests that within the interpersonal realm we are aware of the partiality and the contextuality of self-control and responsibility. It might be said, therefore, that though we have a tendency to assume a libertarian free will we also have a somewhat opposing tendency (implied by our behavior in some circumstances) to be aware of the limitations of self-control and responsibility. So, the image of ‘robust libertarians’ does not suit us, hence, the criticism concerning the horribly confusing effect of EC is exaggerated.

3.1 Reintroduction of the problem of moral luck

There is, however, a further effect of the EC analysis of punishment: It reintroduces the problem of moral luck; namely, it highlights the problem of *uncontrollable interpersonal differences* within the parameters that are pertinent to moral responsibility. These uncontrollable interpersonal differences are in tension with the pursuit of the general social aim of punishment.

I may unpack this argument by an illustration. Consider a distinction between two sorts of cases: Cases of Sort One are those in which a person, who has excellent pertinent psychological abilities and who is not suffering from circumstances which tax her abilities, decides to act immorally and/or to offend the law. In regard to cases of this sort, the EC, assuming the terms of embedded desert, may argue that the lack of control in the ultimate level (i.e., lack of control in the prerequisites that enable one to have good qualities pertinent to moral agency) does not undermine the justification of punishment. However, as mentioned before, embedded desert assumes that the prerequisites that enable one to be a moral agent might be fulfilled to various degrees. As we saw earlier, many factors which the person cannot control may enhance or decrease her qualities as a moral agent; such factors may include the person's level of intelligence, her interpersonal capacities, self-control abilities, various external circumstances and her normative adequacy. As with other human abilities, such as intelligence, we may assume that qualities pertinent to moral agency are distributed normally: Most people have moderate abilities, and there are also smaller numbers of people who are of low or high ability levels. We may call cases in which a person's pertinent qualities are low when compared to the average: Cases of Sort Two. The EC may agree that in cases of Sort Two the agent's desert for punishment is somewhat reduced; namely, in cases of Sort Two there is some injustice from the punished person's perspective, she was disadvantaged from the start. It would be naïve to assume that all or even most of the cases of punishment are pure Sort One cases. So, the EC has to accept that in some (or even most) of the cases of punishment at least some injustice is involved; one need not be an incompatibilist in order to recognize the problem of moral luck in regard to punishment. We may conclude that there is a tension between the general social justification of punishment and the perspective of the punished person. Therefore, we may view the institution of punishment as the lesser of two evils – some injustice seems unavoidable (due to cases of Sort Two) – yet this, as a general social aim, is preferable at the cost of some injustice at the individual level.

One may suggest dealing with this problem by adjusting the punishment to the level of the offender's abilities as a moral agent.[13] Apparently, this may serve as a complete solution. But, as this ‘solution’ requires an assessment in each case of the defender's abilities as a moral

[13] This was pointed out by Mason Cash (in a private correspondence).

agent; the execution of a reliable and valid assessment system seems very speculative (I see a difficulty in finding an objective, applicable and inclusive scale of the abilities pertinent to moral agency; the problem becomes even more complicated by potential distortion by the assessed person, who has an interest in reducing desert). Yet, imprecise mitigation of punishment according to personal circumstances, especially in extreme cases, is possible, but this cannot solve the core of the problem: i.e., the *wide* range of uncontrollable interpersonal differences in the abilities pertinent to moral agency. I therefore believe that these considerations justify rehabilitation aimed at the well-being of the punished person, which may accompany punishment and hence *reduce* (but not eliminate) the injustice involved in it. Thus, I conclude that the EC analysis of punishment suggests that rehabilitation is not a grace but rather an essential ingredient that enables one to justify the institution of punishment. Without it, the legitimacy of the institution of punishment is questionable.

4. Conclusions

Embedded compatibilism embraces a vision. We are prone to have a distorted notion of ourselves as *morally ultimate selves*; this, inter alia, leads to distortions in our moral self-judgment and in moral judgment of other persons. In addition, some of the injustice within our social norms can be related to this notion of the self. Thus, I hope that moral development might be assisted by uncovering this notion of a self which we are prone to have. But as these are dramatic words that do not yet have enough content (and a philosopher should be cautious), may I recapitulate what I have actually done here.

I have argued that we are all morally embedded selves and that this implies that libertarian free will does not apply; neither does pure compatibilism, nor does pure hard determinism. However, we can avoid the state of fundamental dualism (which Smilansky suggests) by creating a revised notion of desert, i.e., *embedded desert*. This analysis supports a hybrid notion of punishment (i.e., one that takes into consideration both the social aim of punishment and desert at the level of the individual agent). In addition, this analysis uncovers the incomplete nature of guilt and reintroduces the problem of moral luck. The problem of moral luck might be reduced, however, by introducing a policy of rehabilitation in addition to punishment. This implies that rehabilitation is not a grace but rather a necessary supplement to punishment.

University of Haifa, Israel
Email: p_guy@013.net.net

REFERENCES

Bomann-Larsen, L., 2010, "Revisionism and Desert", *Criminal Law and Philosophy* **4**, 1–16.

Clark , A., 2007, "Soft Selves and Ecological Control", In *Distributed Cognition and the Will*, Ross, D. Spurrett, D. Kincaid, H. & Stephens, L. G., 101-122, MIT Press, Cambridge.

Double, R., 1991, *The Non-Reality of Free Will*, Oxford University Press, Oxford.

Fischer, J. M. and Ravizza, M., 1998, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press, New York.

Fischer, J. M., 2007, "Compatibilism", In *Four Views on Free Will* (Great Debates in Philosophy), 44-84, co-authored with J. M. Fischer, R. Kane, and D. Pereboom, Wiley-Blackwell, New York.

Hart, H.L.A., 1970, "Punishment and Responsibility", Clarendon Press, Oxford.

Levy, N., 2003, “Cultural Membership and Moral Responsibility”, *The Monist*, **86**, 145-163.

Nagel, T., 1986, *The View from Nowhere*, Oxford University Press, Oxford.

Morris, H., 1968, “Persons and Punishment”, *The Monist*, **52**, 475–501.

Pettit, P., 2007, “Neuroscience and Agent-Control”, In *Distributed Cognition and the Will*, 77-91, Ross, D. Spurrett, D. Kincaid, H. & Stephens, L. G. (eds), MIT Press, Cambridge.

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

Russell, P., 2002, “Critical Notice of Responsibility and Control”, *Canadian Journal of Philosophy*, **32**, 587-606.

Smilansky, S., 2000, *Free Will and Illusion*, Clarendon Press Oxford.

Strawson, P.F., 1974, “Freedom and Resentment”, In *Freedom and Resentment*, Methuen, London (Original work published 1987).

Vargas, M., 2007, “Revisionism”, In *Four Views on Free Will* (Great Debates in Philosophy), 126-165, co-authored with J. M. Fischer, R. Kane, and D. Pereboom, Wiley-Blackwell, New York.

Vargas, M., 2004, “Responsibility and the aims of theory: Strawson and revisionism”, *Pacific Philosophical Quarterly*, **85**, 218-241.

Wolf, S., 2003, “Sanity and the Metaphysics of Responsibility”, In *Free Will*, 372-387, G. Watson (ed), Oxford University Press, Oxford (Original work published 1987).

Wolf, S., 1990, *Freedom within Reason*, Oxford University Press, New York.