CRÍTICA, *Revista Hispanoamericana de Filosofía*
Vol. XXVII, No. 80 (agosto 1995): 119–127

A PARADOX IN COMPATIBILIST ACCOUNTS OF FREE WILL AND MORAL RESPONSIBILITY*

CARLOS MOYA
Universidad de Valencia

Many people are nowadays convinced that compatibilism is adequate as an account of human freedom and moral responsibility. Some people even think that it is the *only* tenable position.[1] It would certainly be nice if compatibilism proved to be such an adequate account, for both libertarian incompatibilism and determinist incompatibilism seem to be too demanding, though in opposite directions. Libertarianism seems to require of us that we accept more than we can understand by means of ordinary science, namely that agents are free as long as they are able to originate actions in a radically self-determining way, i.e., without being subject to the determining role of their beliefs, desires, character, and circumstances, so that an agent 'could have done otherwise' even if all these factors had remained

* Research for this paper has been funded by the Spanish Government's DGICYT as part of the project PB93-1049-C03-02. My thanks to this institution for its generous help and encouragement.

[1] See, for example, A. Ward, 'Talking Sense about Freedom', *Philosophy and Phenomenological Research*, vol. L, no. 4, June 1990. See also G. Watson, 'Free Action and Free Will', *Mind*, 96, 1987, pp. 145–172, esp. p. 169.

unchanged at the time of the action. Libertarians claim that, unless such true origination or self-determination occurs, we are not ultimately justified in believing that we are morally responsible for (some of) our actions.[2] But many people find the price to pay for such true moral responsibility too high: they find unintelligible the 'contra-causal' powers that libertarians postulate. Davidson, for one, invites us to reject the premises of those 'theories that make freedom consist in decisiones taken apart from all desires, habits, and dispositions of the agent'.[3] Determinism, on the other hand, seems to require of us that we abandon more of our everyday view of human agents than we are prepared to and even than we naturally can, namely our firm conviction that we (sometimes) act freely and are morally responsible for (some of) our intentional actions.[4] Determinism shares with libertarians the claim that freedom requires contra-causal origination and that this is a necessary condition for moral responsibility, although, they endorse the truth of determinism, they consequently deny the existence of freedom and moral responsibility. In this situation, compatibilism appears as an attractive middle way: we do not have to give up either causal determinism or everyday belief in freedom and moral responsibility, for the former is compatible with the truth of the latter. Not being uncaused, but having the appropiate sorts of causes, is what distinguishes

[2] This claim is accepted even by some philosophers who deem such self-determination conceptually impossible. See, e.g., Galen Strawson, *Freedom and Belief*, Clarendon Press, Oxford, 1986.

[3] D. Davidson, 'On the Very Idea of a Conceptual Scheme', in *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford, 1984, p. 185.

[4] It is worth pointing out that the intentional character of an action seems to be a necessary condition of its being free. This point is often missed in literature on the freewill problem, which suggests that a closer collaboration with the philosophy of action might be fruitful for philosophical analyses of free will.

free actions and justifies attributions of moral responsibility. Compatibilism presents itself, then, as a reasonable way out between the two apparently unacceptable horns of a dilemma. No wonder it has become so widely accepted. It is a reassuring position to occupy, since otherwise it seems that any new progress in neurological and neurophysiological explanation would be depriving freedom of a further bit of its nourishing soil.

I am afraid, however, that compatibilism is not going to be a comfortable position to stay in, either. In fact, in what I will call its 'classical' form, which roots in Hume's empiricism and finds a paradigmatic formulation in A.J. Ayer's 'Freedom and Necessity',[5] compatibilism has never been a satisfactory account of human freedom and moral responsibility. In identifying free action with unconstrained action, classical compatibilism was never able to explain why we attribute moral responsibility to human beings and not to, say, dogs or cows. Clearly, both human beings and dogs act sometimes free from constraint, both do (sometimes) what they want to do. So, acting freely, in the sense required by moral responsibility, cannot be correctly analysed as acting out of one's own (non-compulsory) desires and beliefs, since otherwise we should hold dogs and cows morally responsible for at least some of their actions. Classical compatibilism, then, did not fulfil its promise of providing a foundation for moral responsibility in a determinist world. Something was presumably missing in classical compatibilist accounts of freedom and moral responsibility, something which human beings possess and which dogs lack.

In fact, Kant had already suggested, two centuries ago, what the missing factor might be: human beings have not

[5] Now in G. Watson (comp.), *Free Will*, Oxford University Press, Oxford, 1982, pp. 15–23.

only desires, but also a rational will, a faculty, distinct from desire, by virtue of which they can be moved to action by the requirements of moral law alone, and so act against their natural desires and appetites. If compatibilism were able to purify this Kantian intuition of its 'contra-causal' flavour and make it coherent with a causal view of human action, then it could considerably improve its own prospects of providing a basis for moral responsibility within a determinist world. The attempt was made by Harry Frankfurt.[6] I will call his analysis of freedom and moral responsibility 'sophisticated compatibilism'.

Let me give just a brief sketch of Frankfurt's theory. As I interpret this theory, the concept in it that corresponds to the Kantian 'will' is that of 'second-order volitions', namely desires about which of our ordinary desires should move us to act. Second-order volitions are a causal, naturalized counterpart to the Kantian 'will'. They are allowed to arise in us as part of a causal chain which includes our heredity, character, environment, and so on, but they are, at the same time, what distinguishes (some) human beings from other animals that have only first-order desires, that is, desires to act in a certain way. Human beings, unlike dogs, are able to reflexively distance themselves from their first-order desires and to adopt positive or negative attitudes towards them from a higher point of view. Being moved by a non-compulsory, first-order desire is acting freely in the sense of classical compatibilism, but it is not enough to have freedom of the will in the sense required by attributions of moral responsibility. Only those beings that can have desires about their first-order desires are able to have freedom of the will and can be morally responsible. And they exert this freedom (they act freely in the higher

[6] See his 'Freedom of the Will and the concept of a Person', ibid., pp. 81–95.

sense) as long as their acts arise from desires by which they (reflexively) want to be moved in their acts. As Frankfurt himself puts it: 'It is in securing the conformity of his will [the desire on which one acts, C.M.] to his second-order volitions, then, that a person exercises freedom of the will.'[7] The concept of (reflexive) identification with one's first-order desires is crucial in Frankfurt's analysis. A free action (an action in which the agent exercises his free will) is one that is caused by a desire with which the agent identifies himself. The compatibilist nature of this account shows itself in that no contra-causal agency is required: the whole process of acting (either freely or not) can be deterministic, with no consequence to the freedom (or lack of it) of the action. Both free and unfree actions are causal, even deterministic processes: the distinction between them lies only in the sorts of causal chains involved in either case. In fact, Frankfurt holds somewhere else that the 'could have done otherwise' requirement is not necessary for moral responsibility.[8]

Sophisticated compatibilism, unlike classical compatibilism, can account for the difference between human beings and other animals. It can also explain why acting out of some non-compulsory desires evinces in us a sense of necessity and unfreedom. And it locates the moral self at a psychological level which does not coincide with that of mere desires to act, getting closer to Kant's distinction between the subject's will and his desires. It has, then considerable theoretical advantages over classical compatibilism.

My central point in this paper, however, is that sophisticated compatibilism faces a paradox that seriously threat-

[7] Ibid., p. 90.

[8] See his well-known paper 'Alternate Possibilities and Moral Responsibility', *Journal of Philosophy*, 66, 1969, pp. 829–839.

ens to undermine its plausibility as an account of human freedom. Here is an outline of the paradox: if sophisticated compatibilism is right, then a person who happens to embrace a deeply disintegrated system of values will enjoy more freedom in his actions than another person whose values show a higher degree of integration. Moreover, the number of his free actions will increase in direct proportion to the degree of disintegration in his values, so that if these are completely disintegrated, then all of his actions will be free. Freedom, then, is obtained at the expense of moral integrity and reliability. This, I think, is an intolerable result. Therefore, the theory that implies it is seriously in trouble. Let me make this point more vivid with an example. Imagine a person who finds altruism highly valuable, but at the same time can see many worthwhile aspects in egoism, so that he embraces both values. So, when he acts so as to promote the welfare of others, he acts freely (his freedom of the will is actually exerted) for he, valuing altruism, identifies himself with the first-order desires that lead him to act in this way. But when he acts so as to foster his own benefit, in spite of the harm done to others, he is acting no less freely, since he, valuing egoism, identifies himself with the corresponding first-order desires. On the contrary, a person who honours altruism, but not egoism, will act freely (will actualize his freedom of the will) in the first case, but not in the second. Now, it is easy to see that the higher the number of pairs of contrary values which that person embraces, the higher the number of his free actions will be. In the end, if he manages to embrace all possible pairs of contrary values, all of his actions will be free, for each will be covered by one or other of those values, that is, each will be caused by a first-order desire with which he identifies. And this is bad news indeed for compatibilism, given that it has to view our odd and morally unreliable agent not only as free, but as the paradigmatic

example of perfect freedom of the will, in that he exercises his free will in all of his acts by 'securing the conformity of his will to his second-order volitions', to use Frankfurt's own words. If we find this consequence unacceptable, we should consider sophisticated compatibilism (not to mention classical compatibilism) as a serious misapprehension of human freedom and moral agency, unless we can find some way out.

I do not see any reason for denying that the consequence actually follows from the theory. So, one can try, perhaps, to suggest that the person in our example cannot possibly exist or to add some correcting clause to the theory in order to dodge the unwanted result. As for the first, one could argue that such a disintegrated system of values cannot possibly be the system an individual can embrace, for it involves contradictions, just as a person cannot be said to have a massively incoherent system of beliefs. I do not think this reply will do, however, because I do not think this person is guilty of any contradiction. He is not holding contradictory views of the good: he is just holding that altruism is valuable for such and such reasons and that egoism is valuable for such and such (different) reasons —and so on for the rest of values—, and he can come to hold these views by means of a perfectly rational process of weighing the pros and cons of taking several values as guides to life. This does not need to involve any contradiction. Through this process, he can become a sort of sincere moral opportunist, applying different moral standards according to his changing desires and circumstances. But if this kind of person cannot be discarded as irrational or contradictory, the second way out, namely adding a clause to the theory in order to avoid the consequence, has an unmistakable *ad hoc* character, not being supported by any other aspect of the theory in any way. Think of adding a clause like 'and the person does not embrace contrary values' to

the conditions of free will stated by sophisticated compatibilism. What has this clause to recommend itself in the context of sophisticated compatibilism except that it is a way to avoid the paradox? Given that this theory does not exclude determinism, our person might be led to embrace this system of values by a causal process, while still retaining all his capacity for judging his own desires from the vantage point of his value-embracing self. And what could this theory reasonably object to the fact that these judgements might be positive in most or even all cases?

I am not claiming that compatibilism will not be able to find a way of dodging the paradox. But if we reflect on the fact that compatibilism, by its very essence, has to conceive of free actions as the result of certain kinds of causes, then it seems to be always possible that a person assemble in himself just that combination of causes which make him, against our best intuitions, into a systematically 'free' agent. In other words, one form or another of the paradox will be threatening any compatibilist attempt to analyse free will, or so it seems.

If we find the paradox really unbearable, and if, moreover, we do not want to stop talking about freedom and moral responsibility, then maybe we should start taking libertarianism more seriously. The paradox seems to arise because compatibilism, even in its sophisticated form, allows the possibility that no real friction between natural desires and moral standards should arise, so that a person can become free by simply ensuring that they are in harmony. The subject in our example does not need to face any conflict between first- and second-order desires, and so has no need to choose between his first-order desires. Moral standards become an expression of desires, even if these are 'second-order' ones, so that there is a chance that no real tension between desire and morality arises. Kant, however, saw clearly that morality and freedom can live only within a

real and insurmountable struggle between practical reason and natural desire, at least in human beings. If this struggle is lessened to one between different level desires, so that it become ultimately eliminable, morality and freedom go by the board too.

Recibido: 20 de marzo de 1995