Philosophical Explorations, Vol. 10, No. 1, March 2007

Routledge
Taylor & Francis Group

REFLECTION, REASON, AND FREE WILL

Timothy Schroeder

Jürgen Habermas has a familiar style of compatibilism to offer, according to which a person has free will insofar as that person responds appropriately to her reasons. But because of the ways in which Habermas understands reasons and causes, he sees a special objection to his style of compatibilism: it is not clear that our reasons can suitably cause our responses. This objection, however, takes us out of the realm of free will and into the realm of mental causation. In this response to Habermas, I focus on the details of his style of compatibilism. I suggest that, while the basic picture is appealing, three key details of it are problematic.

KEYWORDS free will; neuroscience; compatibilism; reason responsiveness

The problem of free will has received so much discussion since the 1970s that it is quite a challenge to say something new on the subject. Nonetheless, Jürgen Habermas has met the challenge in his new paper on free will ('The Language Game of Responsible Agency and the Problem of Free Will'). As I read him, he is inclined toward a currently popular form of compatibilism, according to which one has free will insofar as one's will is guided by one's reasons for action. But he sees a little-discussed problem for this and most other forms of compatibilism: they presuppose that the capacity of the will to cause actions is unproblematic. This, Habermas holds, is far from true. In fact, the really difficult part of a theory of free will is the part where one explains how the will could possibly be understood both as something mental—and so, on his view, essentially normative—and as something within the realm of causes and effects.

In this response, I say little about the struggle for a serviceable theory of mental causation. Instead, I focus on what Habermas says that is specific to his theory of free will. I argue that it suffers from certain weaknesses stemming from its treatment of reflection and reasons. I begin, however, just as Habermas begins: with neuroscience.

1. A Neuroscientific Fantasy

In the minds of many people, if free will is anything then it is a power people have to violate the otherwise lawful order of events in the physical world, and to do so through their own decision-making. Few philosophers writing on free will think of it in these terms, and fewer still believe in such a power. But never mind that for the moment. This is how many scientists think of free will, and many laypeople as well. It is certainly how scientists are thinking of free will when they announce that neuroscience will reveal that

ISSN 1386-9795 print/1741-5918 online/07/010077-8

DOI: 10.1080/13869790601170177

there is no such thing. So take a step back from familiar philosophical presuppositions for a moment. Suppose that a power to violate the physical order by decision-making alone were to exist. What would it be like?

From the perspective of neuroscience, such a power would be fascinating to observe in action. With sufficiently skillful investigative techniques, one could observe neurons in the brain firing without any physical cause, or firing at a rate not explained by the physical influences being exerted. In some limited region of the brain, minor but detectable physical miracles would occur. Already, neuroscience has a good idea of where in the brain to look for such miracles. Assuming the action of the will comes after perception and judgment, and comes before activity in the spinal cord, then neuroscience would be reasonable in expecting to find the will somewhere between the basal ganglia, premotor, and motor cortex (see, e.g., Kandel, Schwartz, and Jessell 2000, chap. 43; Schroeder 2004, chap. 4). If the will were to have the power to cause violations of physical law, then these brain structures would be the place to look to see them. Violations of physical law would show up on carefully designed experiments involving arrays of microelectrodes: somewhere in, say, the globus pallidus of the basal ganglia would be a cluster of neurons contributing to action that would show firing patterns that were partially or wholly independent of the firing patterns of the neurons causally connected to them. There would be many other possible explanations to rule out first, of course. But there is no reason in principle that a will-caused exception to the laws of physics could not be detected within the human brain. There is no reason even to think that we are so far away from having the ability to detect such exceptional activity, were it to exist.

Detection of the miracles would not be the end of it either. The strength of the will (as a physical force) could even be measured. Once the neurons firing out of step with their physical antecedents had been identified, it would be possible to determine where in the firing of an individual neuron the will has its effect. Does it move vesicles of neurotransmitter out to the cell membrane of the axon, causing neurotransmitters to be released into synapses? Does it depolarize the cell membrane of the dendrites? And with these questions answered, it would become possible to create forces to counteract the will. Suppose one attached a very tiny spring, as it were, to a vesicle full of neurotransmitter. Now there is a new question to be answered: how strong could the spring be before the will would be incapable of moving the vesicle? Suppose one artificially created an unusually strong polarizing charge across the dendritic membrane. How strong could the charge be before the will became incapable of causing depolarization? Such tests would reveal the strength of the typical will in standard physical units of force. Perhaps it would even turn out that having a strong will is a matter of having a will capable of exerting six micro-Newtons of force as opposed to a more typical five, or a weak-willed three.

Occasionally people—especially hard-nosed scientists—are inclined to dismiss the notion of a free—free of the laws of physics—will on the grounds that such a thing would be 'unverifiable.' My point so far has simply been to remind the reader not only how easy it would be to verify the existence of such a thing, but how measurable such a thing would be, how easy such a thing would be to assimilate into our existing framework for scientific investigation.

2. A Less Fantastic Neuroscience

The existence of brain events that defy the laws of physics is thus an open, and by some measures serious, epistemic possibility. But it seems unlikely that the objects of

physics will ever be shown to violate the laws of physics. So far as we can tell, the physical order approximately conforms to the physical laws we have so far postulated, and conforms without exception to as yet undiscovered physical laws.

Habermas has offered his readers no reason to doubt that things will turn out otherwise, nor does he wish to. Ultimately he advocates working on a notion of 'top-down,' '"strongly" emergent' causation (p. 40) from the mental to the physical in a way that does not reduce to causation from the physical to the physical. But he seems to accept that any theory of top-down causation needs to be compatible with the 'principle of the conservation of energy' (p. 40). That is, no theory of mental causes in general, and of the causal operation of free will in particular, can be acceptable if it allows mental causes to cause what the laws of physics would regard as miracles. I called the previous section a 'neuroscientific fantasy.' Habermas holds that fantasy is just what it is.

Thus, I suspect that Habermas is not in any fundamental conflict with those scientists, mentioned in his paper (Section 1), who expect science to reveal something about free will in the near future. They expect that science will reveal something about the lawful nature of the physical order within the brain, and Habermas shares their expectations in this regard. Their only disagreement is over whether there is something other than the power to violate this physical order that might deserve the name 'free will.' Habermas holds that there is. (In this respect, at least, Habermas is a compatibilist: the existence of free will is compatible with the non-existence of any psychological power to violate the physical order.) Neither Habermas nor the scientists are inclined to neuroscientific fantasy.

3. Compatibilism and its Discontents

Theories of free will holding that we have free will while rejecting neuroscientific fantasy can fairly be called compatibilist theories, and are found in two main forms. One form holds that free will requires a special metaphysical power, being able to do otherwise than one does, and argues that this is compatible with the inviolability of the physical order. This sort of compatibilism does not interest Habermas. The other main form of compatibilism holds that free will does not require being able to do otherwise than one does. Rather, free will requires the right sort of psychological organization (having the right sorts of beliefs, desires, reasons, and relations between them), and this organization is compatible with the inviolability of the physical order. This is the sort of compatibilism that draws Habermas's attention.

Habermas writes, 'I understand freedom of the will . . . as the mode of how one binds one's own will on the basis of convincing reasons' (p. 19). This characterization of free will requires nothing in particular by way of the ability to do otherwise, but it does require the person with free will to have her psychology organized in a particular way. As a result, Habermas's theory of free will can trace its lineage back to Harry Frankfurt (1971), to similar works from around that time (e.g., Dworkin 1970; Neely 1974; Watson 1975) and perhaps to more recent descendants of these views (e.g., Lehrer 1990; Smith 1991; Stump 1988; Velleman 1992).

Why, then, does Habermas not end his paper at this point? The reason is that Habermas has a foundational worry not addressed by Frankfurt et al. That foundational worry is that shifting between thinking of one's grasp of reasons as, on the one hand, a causal state, and, on the other hand, a semantic state 'is not without consequences' (p. 28). 'For the agent' (p. 28), thinking of oneself as having free will makes sense when one's psychology

is full of premises, but when one thinks of oneself as full of causal processes, the attribution of free will seems to make no sense at all. What is needed is an account of how one's psychology *can* be a cause: an account of mental causation, in other words. Hence Habermas spends the rest of his paper exploring what theory of mental causation one should adopt in the quest to understand how the mental could be both reason and cause.

Habermas holds that the mind is inherently normative: 'propositional attitudes and their contents are part of rule-governed and jointly exercised practices that are oriented towards norms and that can go wrong' (p. 26). And because he holds that minds are normative, he holds that they are not reducible to the non-normative: 'action-motivating reasons are just as irreducible to causally effective events as are the corresponding vocabularies to which these concepts belong' (p. 26). On the same point, Habermas also writes, 'Reasons stand in semantic relations to other reasons. They are fundamentally amenable to critique from opposing reasons, whereas causally explicable mental states or episodes cannot contradict one another' (p. 28). As a naturalistic philosopher of mind who has written with optimism on naturalizing norms, I am not ready to accept these theses as they stand, and I am hardly alone (see, e.g., Dretske 1988; Millikan 1984; Papineau 1987; Searle 2001). However, the issue can hardly be decided here. Habermas is writing within a serious and productive research program (a broadly Wittgensteinian one), and that is a good reason for those of us who do not share his presupposition to set that disagreement to one side, if at all possible, in order to see whether we otherwise agree or disagree with what Habermas thinks about free will in particular. And that is what I will do.

Because of Habermas's views on the nature of the mind, he takes there to be a very deep problem indeed about the nature of mental causation. But this problem is ultimately independent of the problem of free will. Our sense of free will gives the problem a certain urgency, of course, but insofar as we take even non-voluntary mental processes to operate as they do because of the contents of the mental states involved, there is need for an investigation of mental causation regardless of whether we believe in free will or not.

In short, Habermas has a familiar style of compatibilism to offer the reader, but a less familiar sort of objection to it. That objection, however, takes us out of the realm of free will and into the realm of mental causation.

4. Evaluating the Theory

I propose to take for granted that, one way or another, a suitable theory of mental causation will be uncovered. Perhaps it will involve strongly emergent properties and top-down causation, as Habermas suggests. Or perhaps it will not. But however that particular debate gets worked out, once it has been worked out we will have to return to the topic of free will. And once we have returned to the topic of free will, we will be able to ask, is Habermas's proposed theory of free will correct, now that we see how it is possible to have full-blooded mental causation? This is the question to which I now turn.

Recall that Habermas writes, 'I understand freedom of the will . . . as the mode of how one binds one's own will on the basis of convincing reasons' (p. 19). Once the problems of mental causation have been cleared up, how does this theory stand up against its competitors?

To evaluate the theory, it will be helpful to evaluate the principles it is meant to embody. Habermas identifies three. The first is that '[f]reedom depends on the capacity for reflection and self-reflection, the willingness to pause and step back from oneself and

the situation' (p. 16). The second principle is that '[i]n the reflective exercise of free will, the weighing of reasons is linked to the awareness of being able to act otherwise' (p. 16). And the third is that '[s]elf-determination means having the strength of will to ensure that, in acting, one is determined by precisely those reasons that one has found convincing oneself' (p. 16). Habermas's preferred theory of free will comes from uniting these three principles. The first principle is found in the requirement that a free will respond to *reasons*, the second principle is found in the requirement that a free will respond to *convincing* reasons, and the third principle is found in the requirement that a free will act as *self-constraint* on the basis of convincing reasons. Each of these principles is controversial, perhaps false. I will take them in turn.

The first principle holds that free will requires reflective powers and the willingness to use them. It would seem to follow, then, that a person who is unwilling to use his reflective powers in a particular context is unfree, or at least less free than a person who is so willing. Problems lie in wait for such a view. Consider a person engaged in witty banter (I borrow here from Arpaly 2006). Witty banter requires speed, and speed requires foregoing reflection. If one is capable of witty banter, then one listens to what is said, opens one's mouth, and says something witty in response. To reflect is to cease bantering. But now imagine that Dorothy has been challenged to use 'horticulture' in a sentence and she instantly responds 'you can lead a whore to culture but you can't make her think.' Does Dorothy's response lack anything by way of freedom? Does it lack anything by way of moral responsibility? It would seem not, despite the foregone reflection.

Habermas's discussion of his principles has a provision that might seem to help. He indicates that 'a high degree of reflexivity and willpower is only necessary under exceptional circumstances' (p. 18). To be morally responsible it is not necessary always to be exercising paradigmatic free will. Everyday life is full of occasions on which, instead of willingness to exercise reflective powers, one finds 'unclear feelings, dispositions, preferences, and values that direct action pre-reflectively. These motives can be traced to moods, preferences, inclinations, and character traits that often only express traditions, customs, and social norms.' So how is it that we are free (to the extent we are) in acting on the basis of non-reflective habits, moods, and social norms? Habermas holds that 'as long as "we" are the ones to whom the offence is to be attributed, even largely taken-for-granted motives . . . operate with our agreement' (p. 18). This agreement is 'implicit,' he holds.

Implicit agreement is not at all the same thing as actual reflection, however. If one is seeking a theory of the psychological organization of a free will, one will need to decide which it is to be: actual reflection or something that passes for implicit agreement. I see two ways in which one might reduce implicit agreement to explicit reflection, but neither is appealing.

First, one might hold that all implicit agreement to act on certain habits, moods, preferences, and so on stems from earlier reflective decisions. On this view, Dorothy would produce her quip as a result of her current preferences, habits, moods, traits, and so on, but these in turn would be the product of reflective decisions made earlier about whether or not to venture a joke on this occasion, whether to use rough language on that occasion, whether to belittle someone on another occasion, and so on. In this way, Dorothy has not reflectively chosen to say 'you can lead a whore to culture but you can't make her think,' but she has reflectively chosen to become the kind of person who is likely to say such things. This interpretation of Habermas's position is suggested by

the fact that he says we are 'liable for the results of negligent action' when discussing the fact that we sometimes act without reflection.

There are at least two serious problems with this first response. (i) Often, we have no reasonable grounds for predicting what sort of person we will become as a result of present actions. If Dorothy owes her vicious streak in large part to her decision not to flee her difficult home situation at age 16, it is hardly credible that she was aware of the likely ramifications at the time. This is a response Arpaly (2006) has emphasized to such lines of thought. (ii) If Dorothy is only responsible for her present spontaneous acts because they stem from a present character that was formed in the past as the result of free and responsible reflection, then in the past, when Dorothy reflected, her reflection must not have been tainted in any way by unreflective habits, character traits, moods, or the like. But there is no such thing as reflection that is unaffected by non-reflective features of the self. Every act of reflection starts somewhere, and the first thought that begins reflection is not chosen on the basis of reflection. Likewise, which further thoughts enter reflection is not (and cannot be) ultimately based upon reflection. All reflection must rely upon and stem from unreflective processes, or reflection faces a vicious regress from which it cannot escape. And if one is responsible for unreflective actions only because of having engaged in prior free and responsible reflection leading to them, then it seems one cannot be responsible for anything after all.

The second approach to implicit agreement that ties it to reflection is more promising: one could hold that implicit agreement is a matter of counterfactual reflection. If one had reflected, one would have agreed, and this is what implicit agreement amounts to. Dorothy didn't reflect upon whether or not to make her quip, but if she had she would have still made the quip, and this counts as reflection enough.

Here I see a dilemma: is it that Dorothy would *actually* have quipped, had she reflected first? Or is it that Dorothy *should* have seen that quipping was the thing to do, had she reflected? Suppose the former: Dorothy only counts as having quipped out of her own free will if it is true that, had she reflected, she would have still quipped. But this is probably false. Dorothy is in the middle of bantering. Had she actually reflected before saying anything, she would likely have sensed that she had waited too long for any quip to be perceived as witty, and so she would not have quipped, or not in the same way in any case. So suppose it is the latter: Dorothy only counts as having quipped out of her own free will if it is true that she should have seen that quipping was the thing to do, had she reflected. If this is not to succumb to the same objection, this must mean something like 'the facts at the time warranted reaching the conclusion that quipping was the thing to do, and these facts could be appreciated after the fact, in reflection, though not necessarily at the time.' And if this is what is proposed, then it seems that what is important is not the process of reflection itself, but the facts (whatever they are), reflection upon which would show that quipping was the thing to do. These facts, it seems, are what must obtain for Dorothy's quipping to be free, for all that reflection would do is confirm that these facts obtain.

Thus, it seems to me that reflection cannot be held to play any important role in determining whether or not an action expresses one's free will. The first principle should not be accepted.

Turn now to Habermas's second principle. It holds that '[i]n the reflective exercise of free will, the weighing of reasons is linked to the awareness of being able to act otherwise.'

This seems to run afoul of the idea of volitional necessity—the idea that, in some situations, one's reasons are so overwhelming that to do anything other than what one intends to do is unthinkable (see, e.g., Frankfurt 1988, chap. 13). In these situations, it can be the case that reason only permits one action, and one feels this way as well: one is aware that one cannot follow reason in acting otherwise. Oddly, Habermas seems to recognize the possibility of this happening: 'There are hardly ever "knockdown" arguments, usually only arguments that tip the balance' (p. 16). But what hardly ever happens presumably happens on occasion, and what is usual presumably fails to happen in unusual circumstances. But if so, then the second principle can hardly be a necessary condition on the exercise of free will. As many have remarked, if Martin Luther sincerely and correctly uttered the memorable phrase 'Here I stand, I can do no other,' that does not make him any less possessed of free will or moral responsibility than anyone else.

Turn now to the third principle. According to it, '[s]elf-determination means having the strength of will to ensure that, in acting, one is determined by precisely those reasons that one has found convincing oneself' (p. 16). This seems plain enough, but a new family of objections arises. Suppose that one acts contrary to the reasons one has found convincing, but the reasons one acts on are better than the reasons one has found convincing. For example, consider the example of Sam discussed in Arpaly (2003, chap. 2). Sam is a student who needs to get work done. He reasons carefully and is convinced that he should isolate himself from his friends and his usual sources of fun, concentrating exclusively on his studies. The reasons for doing this present themselves to him as very convincing: he needs time to work, non-work activities present possible distractions, and so on. But as it happens, he has reasoned badly. He has forgotten about how depressed and unmotivated he becomes without friends and fun in his life, and how cutting himself off from them has actually slowed his work in the past in similar circumstances. Because of these facts (and because Sam knows of them, even though they do not come to mind), it seems that Sam has better reasons to remain in touch with his friends than he has to cut them off. And now imagine that Sam is actually moved by an unconscious fear of crippling loneliness and a vague sense of foreboding caused by the unconscious action of his memories of the past, moved to give up on his planned isolation. The reasons that Sam finds convincing are still the reasons in favor of cutting himself off from everything but work, but there are clearly better reasons for not doing so, and his appreciation of these reasons—but not any explicit belief that they are better reasons—moves Sam to not cut himself off. Arpaly argues that Sam is more rational in not carrying out his planned isolation than he would have been had he carried out the plan. For my part, I find this quite convincing. But if we agree with Arpaly, and wish to hold to Habermas as well, then sometimes we must say that in exercising a greater degree of free will Sam would act less rationally, and had he acted with less freedom he would have acted more rationally. To my mind, this is a very odd situation to be in. Likewise with moral responsibility: how could some action show one more responsible than the alternative while also showing one to be less rational than the alternative? Thus, the third principle should also be rejected.

For the foregoing reasons, I have some significant hesitations about accepting the principles on which Habermas bases his theory of free will. And because the theory is very much an expression of these principles, I hesitate to accept the theory either. And yet, there are a number of theories of free will now prominent in the literature that emphasize the role of reason's guidance in free will, and these theories are not all beset with the particular problems I have raised for Habermas's own theory (most prominent is

Fischer and Ravizza 1998, which gives a treatment of moral responsibility based on reason-responsiveness). So while I am skeptical about the success of the particular theory just considered, I by no means wish to foreclose the possibility of some similar theory succeeding.

REFERENCES

ARPALY, N. 2003. *Unprincipled virtue: An inquiry into moral agency*. New York: Oxford University Press.

———. 2006. *Merit, meaning, and human bondage: An essay on free will*. Princeton, N.J.: Princeton University Press.

DRETSKE, F. 1988. *Explaining behavior: Reasons in a world of causes*. Cambridge, Mass.: MIT Press.

DWORKIN, G. 1970. Acting freely. *Nous* 4: 367–83.

FISCHER, J., and M. RAVIZZA. 1998. *Responsibility and control: A theory of moral responsibility*. New York: Cambridge University Press.

FRANKFURT, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.

———. 1988. *The importance of what we care about*. New York: Cambridge University Press.

KANDEL, E., J. SCHWARTZ, and T. JESSELL. 2000. *Principles of neural science*. 4th ed. New York: McGraw-Hill.

LEHRER, K. 1990. *Metamind*. New York: Oxford University Press.

MILLIKAN, R. 1984. *Language, thought, and other biological categories*. Cambridge, Mass.: MIT Press.

NEELY, W. 1974. Freedom and desire. *Philosophical Review* 83: 32–54.

PAPINEAU, D. 1987. *Reality and representation*. New York: Blackwell.

SCHROEDER, T. 2004. *Three faces of desire*. New York: Oxford University Press.

SEARLE, J. 2001. *Rationality in action*. Cambridge, Mass.: MIT Press.

SMITH, H. 1991. Varieties of moral worth and moral credit. *Ethics* 101: 279–303.

STUMP, E. 1988. Sanctification and free will. *Journal of Philosophy* 85: 395–420.

VELLEMAN, J. 1992. What happens when someone acts? *Mind* 101: 461–81.

WATSON, G. 1975. Free agency. *Journal of Philosophy* 72: 205–20.

Tim Schroeder, Department of Philosophy, 350 University Hall, 230 North Oval Mall, Columbus, OH 43210-1340, USA. E-mail: schroeder.404@osu.edu