Comments on 'Free Will as Advanced Action Control for Human Social Life and Culture' by Roy F. Baumeister, A. William Crescioni and Jessica L. Alquist

Richard Holton, MIT

I am delighted to be able to comment on this piece by Baumeister, Crescioni and Alquist (henceforth BCA). Baumeister's earlier work has had a huge influence on my own, and I find myself in very substantial agreement with what BCA have to say here.[1] In particular, I agree that if the philosophical debate on free will is to move forward we need to pay close attention to what it is that agents are thinking when they talk of free will, to the experiences that give rise to their conviction that they have free will, and to the effects of such conviction. Moreover, I am in agreement with most of BCA's substantive claims, especially the idea that the experience of free will is somehow tied up with the phenomena of choice and of self-control.

But agreement makes for poor reading. So let me focus on an area where there is disagreement; although even here, what I want to suggest is very much in keeping with BCA's larger project. The issue concerns the relation between deterministic thinking and counterfactual thinking. BCA contend that believing in determinism tends to undermine agents' counterfactual thinking, that is their thinking about what might have been:

> To the lay determinist, everything that happens is inevitable, and nothing else was possible. Thinking about what might have been is thus pointless if not downright absurd, because nothing else might have been (other than what actually happened).

They go on to suggest that such a response is, at the very least consistent with determinism, and perhaps that it is entailed by it:

> The lack of counterfactual thinking in the no-free-will condition can be considered as a straightforward response that is consistent with determinism. After all, if nothing could have happened other than what actually happened, then there are no counterfactuals.

There are two claims that are in play here that need to be separated:

(i) if determinism is true, there are no true counterfactuals;
(ii) if determinism is true counterfactual thinking is pointless.

---

[1] The extent of the influence can be seen in my *Willing, Wanting, Waiting* (Oxford: Clarendon Press, 2009).

It might well be true that ordinary subjects tend to think that these claims are true. Indeed, I will suggest that a tendency along these line explains some of the experimental data that BCA cite. Nevertheless, I think that both claims are clearly false.

To see this, let us start by considering counterfactuals about the past. Imagine a person who, by some stroke of misfortune, finds herself a subject in a psychology experiment. There are various buttons that she can press, some of which bring rewards of varying magnitudes, some of which bring nothing. She observes her own behavior, and the behavior of other subjects around her, and concludes that pushing the red button brings the greatest reward. Looking back on a time when she pressed the green button, she might well make the counterfactual claim:

(1) If I had pressed the red button, I would have got a greater reward

Now suppose that our subject is convinced of the truth of determinism. Should she think that that sentence is false? Surely not. As a result of believing in determinism she might think that at the time she pressed the green button she could not have pressed the red one—or at least, could not have pressed it given the laws and the way the world was prior to her action. But she could still insist that *if* she had pressed the red button, she would have got a greater reward. Such insistence is surely correct. Determinism does not entail that all counterfactuals are false.

Should she hold though that such counterfactual thinking is pointless? Again surely not. If she is a determinist, she will think that the world is governed by causal laws. In particular, she is convinced of the claim that pressing the red button causes her to get a greater reward. But such a causal claim is clearly intimately tied to the counterfactual claim that is expressed by (1). The exact nature of the relation between the causal claims and these counterfactual claims has been a matter of great philosophical debate. But few philosophers now seriously doubt that in some fundamental way counterfactual thinking underlies causal thinking, an idea that has gained much recent support from psychology.[2]

---

[2] The most influential counterfactual account is due to David Lewis, 'Causation', *Journal of Philosophy*, 70 (1973) 556–67, reprinted with post script in his *Philosophical Papers* Vol. II (New York: Oxford University Press, 1986). For subsequent work see John Collins, Edward Hall, and Laurie Paul, (eds.) *Causation and Counterfactuals* (Cambridge, Mass: MIT Press 2004). A recent development understands the counterfactuals in terms of interventions; see Judea Pearl, *Causality* (Cambridge: Cambridge University Press, 2000) and James Woodward *Making Things Happen* (New York: Oxford University Press, 2003) for the account, and the papers collected in Alison Gopnik and Laura Schulz (eds.) *Causal Learning* (New York: Oxford University Press, 2007) for evidence of its psychological applicability.

So let us turn now to what our subject will think about her future action. (1) is a statement about the past. But unless she is very strange she will also accept its present tense analogue:

(2) If I were to press the red button, I would get a greater reward

And she will go on to use this in her reasoning: so, assuming she wants the greater reward, on the next press she will choose red.

Should she somehow be rationally precluded from thinking in this way because she is a determinist? I see no reason that she should be. It is exactly because she is a determinist that she is confident that, once she has understood the relevant causal connections, she can manipulate the outcome. She will of course think that it is determined which button she will press. But that gives her no reason for not pressing the red one.

In particular then, she should be unmoved by an argument that, since it is determined that she will either get the reward or not, there is no point in her pressing the red button. That argument, first formulated by critics of the Stoics, is typically filled out as follows: if it is determined that she will get it, pressing the button is unnecessary; and if it is determined that she will not get it, pressing the button is of no help. So there is no point in pressing it.

The argument doesn't work. Knowing that the outcome is determined gives no reason for thinking that it is determined independently of pressing the red button: that pressing the red button will have no causal influence on the outcome. It gives no reason for thinking that the relevant counterfactual, (2), is false. If our agent knew *which* outcome was determined—if she knew for instance that it was determined that she would get no reward—then she would indeed have no reason to press the red button, since she would know that pressing it would have no effect. Knowing that though, would give her grounds for thinking that (2) is false. In contrast, convinced as she is that (2) is true, she retains every reason to press the red button. It is her faith in determinism that underpins that conviction.

One parenthetical point here is work making, since it might influence one's thinking about counterfactuals. BCA remark that their understanding of counterfactuals might be different to the philosophical one. Their understanding requires, as the term suggests, that counterfactuals are counter to fact—that is, that if a counterfactual is true, then its *if*-condition will not be fulfilled. This is indeed contrary to the philosophical usage, which places no such general requirement; however it is also contrary to normal English usage, as our discussion up till now should reveal. Certainly it would be odd to utter (1)—'If I had pressed the red button then I would have got a greater reward'—if I had indeed pressed the red button. But it would not be at all odd to utter (2)—'If I were to press the red button, I would get a greater reward'— and then to go on to press the red button. Such a statement would be just the kind of thing that an agent would make in deliberating what to do. So any grounds for saying that

counterfactuals must be counter to fact applies at most to their past tense usage. We might do better to replace the term 'counterfactual' with the more neutral 'subjunctive conditional'; but I'll stick to the more normal term, with the understood caveat that they need not be counter to fact.[3]

So much then for the substantial issue of whether determinism is compatible with counterfactual reasoning. What I have not addressed is the further question of whether subjects typically recognize that it is. We have already seen that some philosophers have doubted that it is. The argument that if determinism is true there is no point in doing anything—the so called 'lazy argument'—was made against the Stoics over two thousand years ago, and versions of it have surfaced regularly since.[4] It would be rash to infer from the fact that philosophers have been tempted by a bad argument to the conclusion that ordinary subject are; but it this case I suspect that that may well be true.

The argument involves, in effect, a confusion of determinism with fatalism: determinism is confused with the view, common in much classical myth and elsewhere, that once one's fate is fixed, there is nothing that one can do about it. Of course, if determinism is true there is a sense is which that is right: the course of one's life is fixed before one is born. But in the sense in which the fatalistic claim is relevant to the lazy argument—the sense in which, say, Oedipus was fated to kill his father so that nothing he could do would prevent it—it is in no way entailed by determinism. Fatalism involves the claim that whatever happens a certain outcome will eventuate, and that will typically involve the denial of the very counterfactuals that the determinist will want to affirm.

Nevertheless, if subjects do tend to confuse fatalism with determinism we would expect that to show up in their actions. Here I think that we should look again at the experiments that BCA cite, experiments showing that belief in determinism tends to undermine moral action. That might be because subjects think that morality is incompatible with determinism, and so, thinking that the truth of determinism has shown morality to be an illusion, conclude that there is nothing to stop them behaving badly.

An alternative explanation builds on the idea that subjects tend to confuse determinism with fatalism. If they do, they will come to think that determinism shows

---

[3] Even in their past tense usage there is good reason to say that counterfactuals with true antecedents can be true. It is just that it would normally be misleading to utter them when we can utter something more informative: that I did press the red button and so did get the greater reward. Where we do not yet have this knowledge, it is quite acceptable to utter them. 'If she had pressed the red button she would have got the greater reward' reasons the detective; 'And she did get the greater reward, so I conclude that she did press the red button'.

[4] For a detailed examination of the argument as the Stoics faced it, see Susanne Bobzien, *Determinism and Freedom in Stoic Philosophy* (Oxford: Clarendon Press 1998).

there is nothing they can do to bring about a good outcome, and so they will stop trying to be good. The belief in determinism thus undermines the subjects' self-efficacy. If this is right, there is nothing specifically moral about the effect; one would expect to find the phenomenon manifested in other domains. Belief in determinism should reduce the likelihood that subjects' would stick with a diet, for instance, or would persist with a challenging task.

Of course, such an explanation cries out for an account of why the subjects still behave selfishly. If they really think that their actions will have no effects, why do they bother to do anything at all? Why not simply lie down and do nothing? Interestingly, even the proponents of the lazy argument never tried to persuade determinists that they were committed to that; all they argued was that they should cease to do unpleasant or expensive things in order to achieve long-run goals. Why is that?

Here I think we should turn to Baumeister's earlier work. There he makes an excellent case that human behavior should be understood as the outcome of two quite separate processes. First there are the basic desires and drives; and then, acting to rein them in, comes self-control. Baumeister has convincingly demonstrated that self-control requires real effort; it is, he shows, rather like a muscle.

What happens then when agents' self-efficacy is undermined? It is not that their basic desires and drives are defeated. It is rather, I suggest, than they become skeptical that they will be able to control those desires; and in the face of that skepticism, they fail to apply the effort that is needed even to try. If they were tempted to behave badly, then coming to believe in fatalism makes them less likely to resist that temptation.

If this is right, we have no reason to think that agents are deeply incompatibilist, that is, deeply committed to free will being incompatible with determinism. The incompatibilism is held in place by the confusion of determinism with fatalism, and with the rejection of much counterfactual thinking that goes with that. Remove the confusion—admittedly not an easy task, but surely a possible one—and the incompatibilism will be much less compelling.