METHODE
Analytic Perspectives
www.methodejournal.org

A dialogue on Free Will

Mark Balaguer

California State University
mbalagu@exchange.calstatela.edu

Pages: 1 – 11

Much of the recent discussion concerning the problem of free will has been centered on the compatibilism/incompatibilism dichotomy. Do you think the central role attributed to this dichotomy is well deserved? And, if so, which of the two alternatives is preferable in your opinion?

This question is near and dear to my heart. The second chapter of my book, *Free Will as an Open Scientific Problem*, is almost entirely dedicated to arguing that the compatibilism question does *not* deserve the central role it's been playing in discussions of free will. I definitely think that there's some value to the question, but I think it's essentially irrelevant (except in trivial ways) to questions about the metaphysical nature of human beings and human decision-making processes. To see why I say this, notice first that we can move away from the term 'free will' and focus instead on terms that have clear, stipulated definitions – e.g., 'libertarian-freedom', 'Hume-freedom', 'Frankfurt-freedom,' and so on. This list can be as long as we like; every time someone thinks of a new kind of freedom, we can carefully define it and add it to the list. Now, for most of these kinds of freedom, it's pretty obvious whether or not they're compatible with determinism.

issn: 2481-0498

For instance, if we focus on just the two kinds of freedom that have received the most attention – namely, libertarian-freedom and Hume-freedom – it is more or less obvious that Hume-freedom is compatible with determinism and libertarian-freedom is *not* (indeed, I think that the best definitions of 'libertarian-freedom' build the requirement for indeterminism right into the definition of the term). But given this, I think the question of whether free will is compatible with determinism essentially reduces to the question of what free will *is* – i.e., it boils down to the following:

> *The what-is-free-will question*: Which of the stipulated kinds of freedom – i.e., libertarian-freedom, Hume-freedom, Frankfurt-freedom, etc. – is *real* free will?

For instance, if free will is libertarian-freedom, then free will is not compatible with determinism, but if free will is Hume-freedom or Frankfurt-freedom, then it is. (I'm moving a bit quickly here. Suppose we define *otherwise-freedom* as the ability to make decisions such that you could have done otherwise. Clearly, it's not obvious whether this sort of freedom is compatible with determinism. But this is just because this definition leaves us with more conceptual analysis to do; in particular, we need to say what *ability to do otherwise* is. So rather than saying that the compatibilism question reduces to the question 'What is free will?', a more careful claim would be that the compatibilism question reduces to questions of conceptual analysis – i.e., to questions like, 'What is free will?', 'What is ability to do otherwise?', and so on.)

But what sort of question are we asking when we ask what free will is? Well, the first point to note here is that we're asking a *semantic* question. In particular, the what-is-free-will question is essentially equivalent to the following:

> *Semantic Question (SQ)*: What is the referent of the term 'free will'?; i.e., which of the stipulated kinds of freedom is the referent of 'free will'?

But it's not entirely clear what sort of question this is. To appreciate this, consider the following:

> *Meta-Question*: What kinds of facts determine which answer to SQ is the *right* answer?; i.e., what kinds of facts determine whether 'free will' refers to libertarian-freedom, or Hume-freedom, or Frankfurt-freedom, or…?

This question is surprisingly controversial among those who work on SQ. Some would say that the *right* answer to SQ is the one that captures the ordinary concept of free will; others would give different answers. But I think it can

ISSN: 2281-0498

be argued (and, in fact, I *have* argued this point, in chapter 2 of my book) that no matter how we answer Meta-Question, SQ is not relevant in any non-trivial way to an inquiry into the nature of human decision-making processes. If we're interested in uncovering the metaphysical nature of human decision-making processes, the question we should be focused on (in connection with the topic of free will) is the following:

> *The which-kinds-of-freedom-do-we-have question*: Which of the stipulated kinds of freedom do human beings actually *have*?; i.e., do they have libertarian-freedom?; and do they have Hume-freedom?; and so on.

This is the question about free will that's really about *us*. Now, of course, even if we could answer the which-kinds-of-freedom-do-we-have question, if we couldn't also answer SQ, then we still wouldn't know whether we have *free will*; but that would only be because we lacked some *verbal* knowledge – in particular, knowledge of how the term 'free will' is to be defined. We can put the point like this: we would be able to answer the *metaphysical* part of the do-we-have-free-will question but not the *semantic* part. (This is just a brief characterization of my view on this topic. For a detailed argument for this view – i.e., the view that the compatibilism question and the what-is-free-will question are essentially irrelevant (except in trivial, nit-picky ways) to questions about the metaphysical nature of human decision-making processes – see chapter 2 of my book.).

In the last three decades the discussions on the so-called "Consequence Argument" have convinced many philosophers that compatibilism is not a viable theoretical option. What is your opinion on that argument?

I think the consequence argument gives us some reason to lean toward incompatibilism, but there are also reasons to endorse compatibilism, and the consequence argument doesn't come close to providing a convincing argument that incompatibilism is definitely true. To see why I say this, let P_0 be a complete description of the world at some time in the distant past, before any humans were born; let L be a conjunction of the laws of nature; let P be a complete description of the world at some moment when someone made an ordinary decision that we would ordinarily think of as free (e.g., for specificity, let's suppose that at the relevant moment, Jay decided to order chocolate ice cream rather than vanilla); and let N be an operator such that 'N(X)' means something like 'X is true and no one has ever had any choice about it being true'. Given all of this, the consequence argument can be formulated as follows:

(1) $N(P_0 \wedge L)$.

(2) If $N(P_0 \wedge L)$, and if determinism is true so that $P_0 \wedge L$ entails P, then N(P).

issn: 2281-0498

Therefore,

(3) If determinism is true, then N(P).

This argument is obviously valid, and the conclusion seems to entail incompatibilism, so the only real question here is whether (1) and (2) are true. Now, these two premises might seem immediately obvious, but the problem is that any compatibilist reading of the expressions in N – in particular, 'has a choice about' – will render (1) or (2) false (assuming that determinism is true). Suppose, for instance, that we give a standard Humean conditional analysis of 'has a choice about'; that is, suppose we take 'S had a choice about X being true' to mean something like 'if S had chosen to render X false, then she would have'. Then on the assumption that determinism is true, (2) comes out false; for on the conditional analysis, $N(P_0 \wedge L)$ is true (no one has ever had any choice about $(P_0 \wedge L)$ being true, because even if people chose to render that conjunction false, they wouldn't succeed), and N(P) is false (there was a time when someone had a choice about P being true, for if Jay had chosen to order vanilla ice cream, then he would have, and hence, he would have rendered P false).

On the other hand, if we interpret the expressions in N along standard incompatibilist lines, then (1) and (2) will pretty clearly be true. On an incompatibilist reading of 'has a choice about', it seems pretty clear that none of us has ever had any choice about the past or the laws – i.e., we're not *able* (on an incompatibilist reading of 'able') to alter the past or the laws. And if a complete formulation of the past and the laws entails that Jay is going to order chocolate ice cream, then clearly (on an incompatibilist reading of 'has a choice about') he has no choice about whether he orders chocolate ice cream.

In sum, then, if we give an incompatibilist reading to the expressions in N, then (1) and (2) come out true; and if we give a compatibilist reading to these expressions, then one of those premises will come out false. Moreover, analogous points can be made about alternative formulations of the consequence argument – they go through only if we read the relevant expressions along incompatibilist lines. Thus, if I'm right, then the question of whether the consequence argument succeeds reduces to a question of *meaning*; in particular, it reduces to the question of how we ought to interpret expressions like 'free', 'has a choice about', 'can', 'could have done otherwise', 'able', and so on.

Now, of course, none of this gives us reason to reject the consequence argument. One could grant everything I've said here and still claim that the consequence argument is cogent; for one could claim that (a) premises like (1) and (2) seem intuitively obvious to us, and (b) this is evidence that incompatibilists are right about the meanings of the relevant expressions in those premises. The problem with this, however, is that there are other kinds of cases in which we have compatibilist intuitions. So the incompatibilist intuitions we have in connection

issn: 2281-0498

with the consequence argument are just one set of intuitions among many. So they give us some reason to favor incompatibilism, but they don't come close to providing a compelling case to endorse incompatibilism.

Assuming that libertarianism as such is a viable position, which of the possible libertarian views (such as those centered on agent causation, indeterminist causation or no causation at all) are preferable?

I argued in my book that there's a version of indeterministic event-causal libertarianism that can be defended against all philosophical objections, and I argued that the question of whether this view is true is an open empirical question. According to the version of event-causal libertarianism that I favor, the question of whether we have libertarian-freedom essentially reduces to the question of whether our torn decisions are libertarian-free, where a *torn decision* is a decision in which the person in question (a) has reasons for two or more options and feels torn as to which set of reasons is strongest, i.e., has no conscious belief as to which option is best, given her reasons; and (b) decides without resolving this conflict, so that the person has the experience of "just choosing", or some such thing. Now, in order for a torn decision to be (fully) libertarian-free, it needs to be undetermined in a certain specific way. To zero in on the kind of indeterminacy that's needed here, let's note first that in any torn decision, there will be a set of reasons-based tied-for-best options. E.g., if I'm about to order some ice cream, and if I've ruled out all available favors except for chocolate and vanilla, and if I'm completely torn between those two flavors, then they are the two tied-for-best options. Now, suppose that in this scenario I make a torn decision to order chocolate; i.e., suppose that I "just choose" chocolate. In this sort of case, we can say that the *reasons-based probabilities* of the two options being chosen – or if you'd rather, the *phenomenological probabilities* – are exactly even. Given this, we can define the sort of indeterminacy that's needed for a torn decision to be libertarian-free as follows:

> *TDW-indeterminacy*: In order for a torn decision to be libertarian-free, it needs to be the case that the actual, objective moment-of-choice probabilities of the various tied-for-best options being chosen match the reasons-based probabilities, so that these moment-of-choice probabilities are all roughly even, given the complete state of the world and all the laws of nature, and the choice occurs without any further causal input, i.e., without anything else being significantly causally relevant to which option is chosen.

Now, in order for a decision to be libertarian-free, it's not enough that it be TDW-undetermined; it needs to satisfy other conditions in addition to being undetermined; most importantly, it needs to be *non-random* in a certain agent-involving

issn: 2281-0498

way (e.g., the agent needs to control which option is chosen). But we can ignore these other conditions here; for the purposes of the present question, what matters is the sort of indeterminacy that's needed for libertarian-freedom. For it's easy to see that if a torn decision is undetermined in the above way – i.e., if it's TDW-undetermined – then it will be appropriate to say that which option is chosen is probabilistically caused by events involving the agent's reasons, and so the view I'm describing here is a version of indeterministic event-causal libertarianism. It's important to note, however, that an advocate of this view could also say that there's a non-causal thread in this view; for one might say that what's really going on here is that (a) the agent's reasons deterministically cause her to make a torn decision from among her reasons-based tied-for-best options, and (b) when she goes ahead and chooses one of the tied-for-best options, *nothing* causes the given option to be chosen over the other tied-for-best options. (In other words, you might think that probabilistic causation is just a mixture of deterministic causation and non-causation.)

Finally, you could endorse an agent-causal view by saying this: When a person makes a torn decision, which tied-for-best option is chosen is *caused by the agent* by means of a non-reducible variety of agent causation. But I think the notion of agent causation is unclear and metaphysically suspect, and what's more, I don't think we *gain* anything by appealing to this notion. It might seem that we do, for it might seem that the appeal to agent causation gives us a better way of explaining *why* a given tied-for-best option was chosen. But this isn't a substantive gain because if we ask why the agent agent-caused the choice of the given option, the agent-causal libertarian won't have any better answer than event-causal libertarians have to the original question of why the agent chose the given option.

During the last years, a growing number of philosophers and scientists have advocated sceptical, eliminativist, pessimistic, or illusionistic views on free will. What do you think of these kinds of views?

Well, this depends on the kind of freedom we're talking about. For instance, I think it's pretty obvious that we have *Hume*-freedom, so there's no reason to be pessimistic there. And more generally, for just about all of the compatibilist kinds of freedom that have been articulated in the literature, it's pretty clear that human beings do possess them. However, it is not at all clear that we have libertarian-freedom, and so one might be pessimistic about that kind of freedom. Is this pessimism *warranted*? Well, *some* of the scepticism here is generated by philosophical worries (e.g., the mind argument, the luck objection, and so on), and I think that these arguments all fail. Indeed, in my book, I argued at length for the conclusion that if our torn decisions are undetermined in the manner I described in my answer to question 3 (i.e., if they're TDW-undetermined), then they satisfy all of the other

issn: 2281-0498

conditions needed for libertarian-freedom. In particular, they will be non-random in the appropriate way (i.e., the agent in question will author and control the decision, and so on), and the non-randomness will be appropriately *generated* by the indeterminacy. The argument for this claim is quite long (it takes up a large portion of my book), and I can't summarize it here; but if I'm right about this, then we get the surprising result that the question of whether human beings are libertarian-free reduces to the question of whether our torn decisions are undetermined in the appropriate way (i.e., TDW-undetermined). This is a straightforwardly *empirical* question, and what's more, I argue in my book that it's an open question. In other words, I argue that as of right now, there's no good evidence either way on the question of whether our torn decisions are TDW-undetermined – and, hence, no good evidence either way on the question of whether human beings are libertarian-free. But if this is right, then we should definitely be *worried* that we might not have libertarian-freedom. I don't think there's a good argument for the claim that we definitely *don't* have libertarian-freedom, but there's no good reason to think that we do have it either. Again, given what we currently know, this is an entirely open question.

One might respond to all of this by saying that the question was whether we should be pessimistic about the idea that we have *free will*, not *Hume*-freedom or *libertarian*-freedom. But given what I said in my answer to question 1, it should be pretty clear what I would say to this. Suppose that there are ten different precisely defined kinds of freedom – e.g., Hume-freedom, libertarian-freedom, Frankfurt-freedom, and so on – and suppose that for each of these kinds of freedom, we already knew whether or not human beings possessed it; then we would already have all the relevant metaphysically interesting facts that were *about human beings and human decision-making processes*. We might not know whether human beings had *free will*, but so what? – that would just be because we lacked some *semantic* knowledge, in particular, knowledge of which of the ten kinds of freedom was the referent of the term 'free will'. Thus, once we've discussed the question of whether we have Hume-freedom and libertarian-freedom and so on, I don't think there's anything metaphysically interesting left to discuss in connection with the question of whether we have free will.

A very recent debate concerns the nature of our pre-philosophical views regarding free will. However, some surveys seem to suggest that we naturally tend towards compatibilism, others that we naturally tend towards incompatibilism. What do you think is the value of this kind of "experimental philosophy" in regard to the issue of free will?

As long as the methodology of experimental philosophy is limited to questions like "What is free will?," I think it is a legitimate methodology. Let's say that a

issn: 2281-0498

type-C question is a question of the form, "What is C?", or "What is a C?", where C is some concept, like *free will*, or *knowledge*, or *person*, or whatever. As I made clear in my response to question 1, I think these questions are semantic questions. Moreover, I think it can be argued that these questions should be interpreted as questions about ordinary language – about what ordinary folk mean by the relevant terms. This is of course controversial, but if it's right, then the intuitions of ordinary folk can be used as data points to confirm and falsify theories, and so the methodology of "ex-phi" is legitimate.

But three cautionary points are in order. First, if the methodology of experimental philosophy is applied to anything but a type-C question, then it's entirely illegitimate. For instance, if we applied it to the question, "Do humans have libertarian free will?," the results would be entirely worthless because there is no reason to think that the intuitions of the folk track the facts in connection with questions like this. The question of whether we have libertarian free will is a non-semantic about-the-world question; it's analogous to a question like "Does Alpha Centauri have planets?" We can't figure out whether Alpha Centauri has planets by polling the folk, and likewise, we can't figure out whether humans have libertarian free will by polling the folk. But type-C questions are different; they are essentially *semantic* questions, and so (assuming a certain metaphilosophical view of semantic questions like this) the intuitions of the folk are relevant. In short, they're relevant here for the same reason that they're relevant in linguistics – because the questions at issue are questions about ordinary language.

Second cautionary point: I think it's extremely difficult to set up good studies in experimental philosophy. One reason for this is that it takes a bit of philosophical training to learn how to play the intuitions-about-thought-experiments game; i.e., it takes some training to learn how you're supposed to respond. As a result, I think that the subjects in these experiments are often not answering according to their own intuitions; they're guessing what the "right" answer is, or what the experimenter wants them to say, or something like that. Moreover, there's something of a catch-22 here because once a person has received the relevant sort of philosophical training, their intuitions are no longer 100% trustworthy because it's possible that their semantic intentions (or what they mean by their words) has been *warped* by the philosophical training. Once someone has come to believe that free will is, say, Hume-freedom, then their intuitions will very likely line up with this; but that doesn't mean that this is how their intuitions would have gone before they came to accept the Humean theory.

Third cautionary point: I think that most of us have compatibilist intuitions in connection with some cases and incompatibilist intuitions in connection with others. I think that the reason for this is that the folk concept of free will is something of a mess – i.e., it's not a precisely defined concept. In particular, I think that

issn: 2281-0498

our concept of free will is imprecise in a way that makes it the case that there is no uniquely correct answer to the what-is-free-will question; indeed, I think our concept is consistent with compatibilist precisifications and incompatibilist precisifications, so that there's no determinate fact of the matter as to whether free will is compatible with determinism. But if all of this is right, then when we do ex-phi, we should refrain from assuming that there are uniquely correct answers to the what-is-free-will question and the compatibilism question. Any given study showing that people have compatibilist or incompatibilist intuitions about some given scenario should be taken with a grain of salt, for they don't establish any sweeping claims about the truth of compatibilism or incompatibilism.

What do you think the relationship is between free will and moral responsibility? With regard to this, do you think that the famous Frankfurt scenarios are crucial for assessing the issue?

I don't have anything very original to say about Frankfurt scenarios, so I will focus on the first part of the question, the part about the relationship between free will and moral responsibility. In my answers to previous questions, I've been making heavy use of the idea that there are multiple kinds of freedom – e.g., Hume-freedom, libertarian-freedom, and so on. I think that the same thing can be said about moral responsibility. Even if we pretend that aside from the freedom requirement for moral responsibility, everything else about the concept of moral responsibility is clear, we can still say that there are multiple kinds of moral responsibility that require different kinds of free will. For instance, there's a kind of moral responsibility that requires libertarian-freedom; there's a kind that requires Hume-freedom; and so on. Now, of course, there's also a kind of responsibility that doesn't require *any* kind of free will, but this isn't a very interesting kind of responsibility – it doesn't seem to correspond to anything that any of us have in mind in connection with the concept of moral responsibility. So it seems to me that we can definitely say this: Moral responsibility (i.e., *real* moral responsibility) requires *some* kind of freedom. Moreover, it seems safe to say that at the very least, it requires some *compatibilist* kind of freedom. Whether it also requires libertarian-freedom is another matter. This is obviously controversial. But what does this question *turn* on? Well, on my view, it turns on whether the term 'moral responsibility' refers to libertarian-responsibility or some other kind of responsibility, e.g., Hume-responsibility. In other words, the question of whether moral responsibility requires libertarian-freedom turns on the answer to the following question:

> *SQ2*: Which of the precisely defined kinds of moral responsibility is the referent of the term 'moral responsibility'?

This, of course, is a semantic question. And according to the metaphilosophical

issn: 2281-0498

view that I endorse, it's ultimately about the folk concept of moral responsibility. Moreover, as is the case with the folk concept of free will (see my answer to question 5), it seems to me that the folk concept of moral responsibility is something of a mess – i.e., it's not precisely defined. In particular, I think that in some scenarios, we have the intuition that moral responsibility requires libertarian free will, whereas in other scenarios, we have the intuition that it doesn't require libertarian free will. So in the end, I don't think there's any uniquely correct answer to the question of whether moral responsibility requires libertarian free will.

Finally, the question of whether moral responsibility requires *free will* (as opposed to Hume-freedom or libertarian-freedom or whatever) boils down to this question: Is the kind of freedom that corresponds to the folk notion of free will required for the kind of moral responsibility that corresponds to the folk notion of moral responsibility? In response to this question, my inclination is to say that the folk notions of free will and moral responsibility are imprecise in parallel ways. And if this is right, then we can say that moral responsibility *does* require free will.

Given the evidence coming from neuroscience and genetics, during the last few years a growing number of scholars have been arguing that the idea that we deserve blame for our bad deeds (and punishment for the worst of them) is ungrounded and should be abandoned. What is your opinion of this view?

Well, if blameworthiness requires only a compatibilistic kind of freedom, like Hume-freedom or Frankfurt-freedom, then the arguments in question here are presumably not cogent. In other words, it's pretty obvious that there's no evidence from neuroscience or genetics that we don't have Hume-freedom. But let's assume that blameworthiness requires *libertarian*-freedom. Given this, the question here is whether there is any good evidence from neuroscience or genetics for thinking that we don't have libertarian-freedom. Let me say a bit about this question.

First, given the view of libertarian-freedom that I articulated in my answers to questions 3 and 4, it's pretty clear that there's no evidence from *genetics* that we don't have libertarian-freedom. On the view I described there (and more fully in my book), we have libertarian-freedom if and only if some of our torn decisions are TDW-undetermined. But even if our genes predispose us to behave in certain ways, there's no good reason to think that they causally determine all of our torn decisions; the claim that at least some of our torn decisions are TDW-undetermined seems to me to be obviously compatible with everything we know about human genetics.

But neuroscience is a different story. *Prima facie*, it doesn't seem unreasonable to think that all of our torn decisions might be determined by prior neural events – i.e., by neural events that occur just before our torn decisions. So the question we need to ask is this: Does neuroscience give us any good evidence for thinking that

issn: 2281-0498

our torn decisions actually *are* determined by prior neural events? The answer, I think, is that it doesn't; in other words, neuroscience does *not* give us any good reason to think that all of our torn decisions are causally determined. Two points are in order here.

First, since torn decisions are presumably *themselves* neural events, you might think that they're determined by prior neural events for the simple reason that all neural events are determined by prior neural events. But, in fact, there is no good reason to believe that all neural events are causally determined. As I point out in chapter 4 of my book, neuroscience is a probabilistic science. Various neural processes (e.g., synaptic transmission, spike firing, and the opening and closing of ion channels) are treated probabilistically by current neuroscience. Now, it doesn't follow from this that these processes are genuinely indeterministic. But the point is that as of right now, there is no good reason to think that they *aren't* indeterministic. It's simply an open question. So if there's any reason to think that our torn decisions are determined by prior neural events (or at any rate, that they aren't TDW-undetermined), these reasons would have to be specifically concerned with torn decisions in particular. Now, you might think that there are some neuroscientific studies out there that show that our torn decisions aren't undetermined in the way that's needed for libertarian free will (i.e., that they aren't TDW-undetermined). In particular, you might think this point is established by the studies performed by people like Benjamin Libet. This, however, leads me to the second point I want to make here: I don't think that the anti-free-will arguments based on these studies are cogent. I don't have the space here to explain what's wrong with these arguments, but I discuss this in chapter 4 of my book.

issn: 2281-0498