Penultimate Draft: Forthcoming in *Experimental Philosophy and Metaphysics*, David Rose (ed.), Bloomsbury.

The folk psychological roots of free will

Joshua Shepherd

[A] 1. Introduction

Debates surrounding free will are notorious for their intractability. This is so in spite of the fact that, even at a fairly fine grain of analysis, competing views on the nature of free will are well understood. Why can't philosophers find common ground? One line of thought that has emerged fairly recently draws on the psychology of concepts. The general idea is that an explanation for persistent disagreement about free will, and perhaps guidance towards resolution, might be found by exploring the psychological roots of 'our concept' of free will – e.g., those psychological factors that underlie our tendencies to say, of some bit of human behavior, that it was performed of an agent's own free will, or not.

This very general idea has motivated very different proposals regarding the psychological roots of free will. I mention two examples. Shaun Nichols and Joshua Knobe (2007) appeal to a difference between responses to cases described abstractly and cases described concretely to argue that the appearance of compatibilist tendencies in applications of the concept free will – that is, attribution of free will to agents in deterministic universes – in fact represents an error. Concrete cases influence application tendencies by stimulating an affective response that biases the judgments. Unbiased applications of the concept, so goes the thought, are consistent with an incompatibilism between free will and determinism. In stark contrast, Dylan

Murray and Eddy Nahmias (2014) use mediation analysis over a range of concept applications (including free will and moral responsibility) to argue that the appearance of incompatibilist tendencies in fact represents an error. Instead of affirming incompatiblism as philosophers understand it, many people misinterpret determinism as implying the bypassing of the normal causal role played by an agent's conscious mental states. Scrubbed of this mistake, the conceptual applications of most participants in experiments on free will are consistent with compatibilism.[i]

Notice that these proposals can be taken to differ not only with respect to the descriptive facts regarding the psychological processes underlying free will judgments, but with respect to the more normative debate regarding what we might call the proper concept of free will. What both proposals share is the aim of using data about our concept to influence the traditional debate and the proper concept. To make good on this aim, it seems we need answers to two questions. First, what are the psychological roots of our concept of free will? Second, how might progress on the first question contribute to progress regarding normative debates about the proper concept of free will?

In sections two and three I address the first question. Section two discusses recent work in the experimental philosophy of free will, and motivates the study I report in section three. Section four reflects on the second question in light of the reported results. To preview, the results suggest that the psychological structure of our concept of free will is sensitive to three independent features: Liberty, Ensurance, and Consciousness. I argue this supports the view that our concept is incompatibilist more than the view that our concept is compatibilist, and I discuss two proposals regarding the normative upshot. On one proposal, these results might be taken to offer some

support to incompatibilism about the proper concept. A second proposal, however, makes room for a much different upshot.

[A] 2. Recent experimental philosophy of free will

In a recent paper, Joshua May (2014) argues that the 'ordinary' concept of free will possesses 'non-classical' structure, in the sense that application conditions for this concept are not explicable in terms of necessary and sufficient conditions.[ii] Rather, application of free will is governed in a graded manner by unrelated features. May proposes two: what he calls Liberty and Ensurance.

According to May, 'an agent has Liberty in a situation just when she has at least two genuine options for action in that situation' (2851). And 'an agent has Ensurance with respect to an action just when the action depends in an appropriate way on her mental states and her environment' (2851). These actual definitions are less important than the motivation behind positing these two general features. As May notes, Liberty is intuitively related to the incompatibilist insistence on the importance of indeterminism in the causal stream leading to free action. And Ensurance is intuitively related to the compatibilist insistence on the importance of control with respect to the causal stream leading to free action. On May's proposal, both features play an important role in normal attributions of free will.

> When both [Liberty and Ensurance] appear to be present, free will is judged to be present. When both factors appear to be absent, free will is judged to be absent. When one but not the other factor is present, it is unclear whether or not free will is present. (2851)

In order to test this proposal, May designed vignettes that varied the presence or absence of Liberty and Ensurance. Liberty was varied via the following paragraph.

> Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature [needn't/will] cause the exact same events for the entire history of the universe. So, every time the universe is re-created, [some things may not/everything will] happen the exact same way.

Ensurance was varied via a story regarding an agent who deliberates and decides to steal a necklace. Normal deliberation and decision constituted the presence of Ensurance. The lack of Ensurance was emphasized via a causal route to action that involved brainwashing. When the agent lacked Ensurance, brainwashing gave her a powerful urge to steal the necklace, and the urge leads to the theft.

One might worry whether May's vignettes really measure the lack of Ensurance. Those vignettes introduce brainwashing, which may generate intuitions related to manipulation rather than simply the lack of Ensurance. May is aware of this problem, and argues that 'attributions of free will are undermined by manipulation via lack of Ensurance' (2860). This might be right. One way to make sure is to run vignettes that operationalize lack of Ensurance in some other way. I do so in the study reported in section three.

As for May's study, the results support his proposal. Both Liberty and Ensurance significantly impacted free will attributions, such that free will attribution was at its highest when both factors were present, and lowest when both were absent.

This is an interesting and important result. It indicates that application of free will is driven by features connected to the traditional dispute between incompatibilists and compatibilists. One might wonder, however, whether these are the only two relevant features. In connection with this question, further recent experimental work begs for attention.

In recent experimental work of my own I have offered evidence that consciousness is, in certain respects, crucial to free will attribution. Consider, for example, study three from Shepherd (2012). In this study participants read vignettes that varied the presence or absence of determinism, as well as the presence or absence of consciousness in the causal stream leading to behavior. Both factors strongly influenced free will attributions, suggesting that conscious causation of behavior is as important for the normal view of free will as is indeterminism.

Study three in Shepherd (2015) goes further than this. In this study participants read vignettes about 'humanoid machines' that behave indistinguishably from human beings. In one vignette these humanoids were described as conscious – the humanoids were said to 'actually feel pain, experience emotions, see colors, and consciously deliberate about what to do' (939). In a contrasting vignette the humanoids were said to lack consciousness. Although the *behavior* of both types of humanoids was described identically, the presence or absence of consciousness made a large difference for free will attributions. The conscious humanoid was deemed to have free will. The non-conscious humanoid was deemed to lack it.

These studies suggest that in addition to Liberty and Ensurance, consciousness may be a third feature important for normal attributions of free will. Alternatively, it could be that consciousness is important for free will, but that its importance is largely

indirect. Perhaps, that is, consciousness is important for free will because it is important for Liberty or Ensurance.

In section three I report a study that tests these possibilities.

[A] 3. Liberty, Ensurance, Consciousness: A study

We are interested in the relationship between consciousness, Liberty, and Ensurance as they relate to free will attribution. In order to test these relationships, I devised a study that independently varied all three factors (see appendix for the vignettes). This design allows, first, for a conceptual replication of May's results. If Liberty and Ensurance are independently important for free will attributions, then we should see independent effects for both, as May did. Second, this design affords a test of various ways consciousness may be important for free will attribution.

Consciousness may impact free will attribution independently, much as May's results suggest Liberty and Ensurance do. If so, we should expect to see an independent effect of consciousness in the following study. Conversely, consciousness may impact free will attribution by tapping into deeper features within our concept of free will – e.g., by tapping into Liberty or Ensurance. If so, we should expect to see interactions between consciousness and Liberty and/or consciousness and Ensurance. Of course, these possibilities are not exclusive. We might find an independent effect for consciousness, as well as an interaction between consciousness and Liberty and/or Ensurance.

[B] 3.1 Participants

520 participants saw one of eight vignettes. Participants were recruited via Mechanical Turk and were paid $.45 for roughly 4-5 minutes of time. Participants

who failed a comprehension question or who failed to complete the survey were excluded from analysis. After exclusion, 456 participants remained (mean age = 34.9, gender = 57.2% male).

[B] 3.2 Design

Participants first read a paragraph, drawn from May's study, that emphasized Liberty or its absence:

> Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature [needn't/will] cause the exact same events for the entire history of the universe. So, every time the universe is re-created, [some things may not/everything will] happen the exact same way.

They next read a paragraph that emphasized Consciousness or its absence:

> Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several psychologists and they assure him that the problem has to do with mechanisms of [non]conscious impulse control in his brain, mechanisms of which [he is consciously aware/he has no awareness]. If Mr. Q's [non]conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Finally, participants read a paragraph that raised the possibility of Ensurance, and then specified whether Ensurance was present or absent. When Ensurance was present, the agent did what she really wanted to do, and refrained from uttering an obscenity. When Ensurance was absent, the agent uttered the obscenity. For example, here is a paragraph that holds the presence of Liberty and consciousness constant, and emphasizes Ensurance:

> Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives in Universe 32, the initial conditions of the universe and the laws of nature do not determine what he will do. Two, there is the issue of whether Mr. Q's conscious impulse control mechanisms are focused in the right way. As it happens, Mr. Q's impulse control is consciously focused, and he refrains from uttering the obscenity.

Participants then were asked to agree or disagree with the statement 'Mr. Q [refrains/utters the obscenity] of his own free will.' Answers were rated on a 1-7 scale, where 1 represented 'strongly disagree,' 2 'disagree,' 3 'somewhat disagree,' 4 'neither agree nor disagree,' 5 'somewhat agree,' 6 'agree,' and 7 'strongly agree.' Finally, participants were asked what factors were relevant to their answer, and were given four options:

> (1) The fact that Mr. Q lives in Universe 32.

(2) The fact that Mr. Q's impulse control mechanisms were [not] [non]consciously focused in the right way.

(3) Both of the above factors.

(4) None of the above factors.

[B] 3.3 Results

A 2x2x2 Analysis of Variance test revealed significant main effects for Liberty, $F(1,456) = 63.911$, $p<.001$, partial eta squared=.125, for Ensurance, $F(1,456) = 19.004$, $p<.001$, partial eta squared=.041, and for Consciousness, $F(1,456) = 23.164$, partial eta squared=.049. There were no significant interactions, although an interaction between Consciousness and Ensurance approached significance ($p=.079$).

The means for each vignette are represented in Table 1.

Table 1. Means by vignette.

Liberty+/Ensurance+/Consciousness+	M = 5.51, SD = 1.29 (N=55)
Liberty+/Ensurance+/Consciousness-	M = 4.63, SD = 1.69 (N=54)
Liberty+/Ensurance-/Consciousness+	M = 4.76, SD = 1.57 (N=54)
Liberty+/Ensurance-/Consciousness-	M = 4.45, SD = 1.61 (N=64)
Liberty-/Ensurance+/Consciousness+	M = 4.61, SD = 1.71 (N=57)
Liberty-/Ensurance+/Consciousness-	M = 3.49, SD = 1.70

	(N=59)
Liberty-/Ensurance-/Consciousness+	M = 3.50, SD = 1.80 (N=56)
Liberty-/Ensurance-/Consciousness-	M = 2.88, SD = 1.57 (N=57)

Regarding which factors were reported as relevant to free will attributions, I computed the percentage of responses by case. These are represented in Table 2.

Table 2. Reported reasons for judgment by vignette. Factor 1 = [In]determinism; Factor 2 = [Non]Consciousness; Factor 3 = Both Factors 1 and 2; Factor 4 = Neither Factors 1, 2, nor 3.

Liberty+/Ensurance+/Consciousness+	1 = 3.6%, 2 = 58.2%, 3 = 34.5%, 4 = 3.6%
Liberty+/Ensurance+/Consciousness-	1 = 7.4%, 2 = 53.7%, 3 = 33.3%, 4 = 5.6%
Liberty+/Ensurance-/Consciousness+	1 = 13.0%, 2 = 48.1%, 3 = 35.2%, 4 = 3.7%
Liberty+/Ensurance-/Consciousness-	1 = 3.1%, 2 = 56.3%, 3 = 34.4%, 4 = 6.3%
Liberty-/Ensurance+/Consciousness+	1 = 22.8%, 2 = 31.6%, 3 = 42.1%, 4 = 3.5%
Liberty-/Ensurance+/Consciousness-	1 = 22.0%, 2 = 18.6%, 3 = 57.6%, 4 = 1.7%
Liberty-/Ensurance-/Consciousness+	1 = 25.0%, 2 = 21.4%, 3 = 51.8%, 4 =

	1.8%
Liberty-/Ensurance-/Consciousness-	1 = 28.1%, 2 = 15.8%, 3 = 56.1%, 4 = 0.0%

Visual inspection of the data suggests an interesting shift with respect to Liberty. When Liberty is present, participants rarely cited the fact that the agent lived in an indeterministic universe as relevant to their free will attribution. But when Liberty was absent, this changed. A post-hoc chi squared test confirmed that the influence of Liberty on which factor was selected was statistically significant X^2 (3, N = 456) = 66.27, p<.001. Neither Ensurance nor Consciousness significantly impacted selection (ps>.75).

[B] 3.4 Discussion

These results offer a conceptual replication of those reported by May (2014), while avoiding worries about manipulation that might problematize his results. Both Liberty and Ensurance are relevant to free will attribution. In addition, these results go beyond those reported by May, by demonstrating that Consciousness is a third, independent factor relevant to free will attribution. In general, when all three of these features are present, free will is attributed at a high level. Take any one of these features away, and answers reflect some uncertainty about the presence of free will. If two features are absent, free will attribution tends to be withheld. If all three are absent, this tendency is further increased.

Recall that May put his results to work in support of an explanation for the intractability of the traditional debate between compatibilists and incompatibilists. May wrote:

> The cluster concept account is meant to help explain the long-standing debate about free will and determinism. It is undoubtedly bold to do this, and I accordingly wish to tread lightly. Nevertheless, we arguably have good empirical and 'armchair' reasons for the idea that the concept of free will is not associated with a single feature that is either compatible with determinism or not. Rather Ensurance and Liberty both play an important role in the concept of free will. (2865)

The co-importance of Liberty and Ensurance certainly *suggests* an explanation. Filling that explanation out, however, requires far more than these results show.[iii] That is to say, I find it plausible that there is some intuitive conflict in our very notion of free will. Indeed, when given the right kinds of cases, I can feel the conflict within myself. But certainly many other factors – arguably involving the influence of particular philosophers and particular arguments, as well as intellectual currents within philosophy and within the culture more broadly at any given historical moment – have contributed to the nature and structure of free will debates as they have developed over time.

I turn to a different issue. The fact that application of the concept of free will is driven by Ensurance, Liberty and Consciousness might be taken to conflict with a recent theory of free will attribution. Murray and Nahmias (2014) argue that most people lack intuitions in favor of incompatibilism, and that attributions of free will that appear incompatibilist can be explained by way of an error theory. According to Murray and Nahmias, non-philosophers misinterpret descriptions of determinism as implying bypassing of those (conscious) mental states, events and processes normally

thought to be relevant for free action. When reading vignettes regarding action in deterministic universes, non-philosophers falsely infer that an agent's mental states, events, and processes 'have no causal effect' on their behavior (2014, 440).

If Murray and Nahmias are right, one would predict different results than those found here. First, one would predict that Liberty would not have an effect independent of that produced by Consciousness. For once the causal importance of Consciousness is properly emphasized, one would think that most participants would not be inclined to mistake determinism for bypassing. Second, one would predict that participants attributing low levels of free will in deterministic, non-conscious scenarios would tend to blame the absence of Consciousness, rather than the absence of Liberty. But this is not what we find. Instead, we find that participants cite Liberty as important to their judgment at significantly higher rates when cases involve determinism, even when these cases also emphasize the absence of consciousness.

There is a further problem for Murray and Nahmias. Although their bypassing hypothesis is strictly silent on the factors influencing free will attribution in indeterministic cases, their overarching aim is to argue that the folk are compatibilists. If so, one would expect that a high percentage of participants who see indeterministic cases would cite Consciousness, but not Liberty, as the reason for their attribution of free will. While a high percentage of participants did cite Consciousness in such cases, over 40% cited Liberty as among their reasons. So a substantial percentage of participants offer justifications that are explicitly incompatibillist. And those that did not offer justifications that are consistent with incompatibilism.

In connection with this last point, philosophical libertarians – those who affirm an incompatibillism between free will and determinism, and assert that we have free will – may wish to press the following point. As May and I have introduced

it, Ensurance is supposed to closely track features that motivate compatibilism. But is this correct? May's operationalization of Ensurance involved action success when Ensurance was present, and brainwashing when Ensurance was absent. With respect to this operationalization, libertarians are likely to complain that their best theories rule out free will in brainwashing cases as well. So May's results regarding Ensurance support folk libertarianism as much as folk compatibillism. My operationalization of Ensurance was different, contrasting action caused in part by effective impulse control in one case with action caused in part by ineffective impulse control in another. Again, however, many libertarians will agree that an agent is not free, or is perhaps less free, when her control over her action is in some way compromised. This is because indeterminism is merely a necessary condition on free will. As Widerker and Schnall recently put the point:

> For a libertarian, indeterminism per se neither undermines free agency, nor renders an action free. Rather, it is a condition that makes the performance of free action possible, a condition that makes room, or provides the opportunity, for the exercise of free agency. (2015, 94)

These are fair points. In my view, the results reported above support the claim that the folk are incompatibilists more than the claim that the folk are compatibilists. Whether the support is strong enough to move the normative needle regarding the structure of the proper concept of free will is less transparent. For the relationship between information regarding the psychological structure of our concept of free will, and the view we ought to adopt regarding the proper concept of free will, is complex. In response to the libertarian, the compatibilist may well reemphasize some of the

many arguments against the cogency of any incompatibilist view. One such argument appeals to the need for an absence of luck in free action. The enterprising compatibilist might make a case that the importance of Ensurance – as I have operationalized it – provides an intuitive strand in our concept of free will upon which absence of luck arguments can build. For the enterprising compatibilist, then, the proper concept of free will may turn out to be compatibilist, once we have cleansed the folk concept of its undesirable incompatibilist elements.

Here I am simply pointing out the existence of dialectical strategies left open by the results reported here. More specificity regarding how we ought to proceed requires greater clarity on the question I elucidated above. How might progress in our understanding of the psychological roots of (our concept of) free will contribute to progress regarding normative debates about the proper concept of free will? In the next section, I offer an appraisal of answers to this question.

[A] 4. Putting x-phi to metaphysical work

X-phi was once – for some, probably still is – associated with the striking image of a burning armchair. Perhaps more than any particular argument, this image gave philosophers the impression that x-phi's philosophical value had to do with the destruction of a kind of orthodoxy. That orthodoxy sees philosophy as centrally (even if not exclusively) concerned with the clarification and analysis of concepts, and as centrally (even if not exclusively) committed to a particular methodology for doing so. This methodology is sometimes called the method of cases, and consists in part of considering a wide range of thought experiments, offering intuitions or judgments about these cases, and thereby collecting a kind of data to which philosophical theories and analyses must in some way answer.

It is no longer clear that this is, or ever was, the best way to understand x-phi's philosophical value. It is true that proponents of the so-called negative program (Alexander et al. 2010) seek to undermine a kind of over-reliance on intuitions. The influence of that program has, in my view, been largely salutary. For example, philosophers are now more self-critical when deploying the method of cases, and more aware of psychological factors that can lead to biased judgments (e.g., order effects). But beyond the negative program, recent reflection on the positive value of x-phi has generated more than one interesting proposal. In the remainder of this section, I consider two such proposals, and reflect on the ways such proposals might illuminate the use of data regarding free will judgments in debates about the proper concept of free will.

One proposal is due to Uriah Kriegel, writing in this volume. Kriegel argues that x-phi can and should contribute to the kind of philosophical program Frank Jackson calls 'serious metaphysics.' As Jackson has it, a major aim of philosophy is the development of a total theory of the world in terms of fundamental and derivative statements and notions, and in terms of entailment relations between fundamental and derivative statements and notions. Conceptual analysis plays a crucial role in the development of such a theory, since many of the relevant statements, notions, and entailments go well beyond any formal structure the theory will possess.

On Kriegel's proposal, the kind of conceptual analysis needed elucidates the meaning of a concept by mapping platitudes about the concept (drawn from commonsense and/or scientific inquiry) into variations of the Ramsey sentence. The variation called for may depend upon particulars of the concept. In response to criticisms of this methodology, Kriegel offers a few alternatives. One interesting alternative, offered in response to Quine's rejection of the analytic/synthetic

distinction, Kriegel calls this the Lewis sentence. A concept analyzed via a Lewis sentence need not contain any nonnegotiable platitude, for a Lewis sentence consists of a disjunction of conjunctions of *most* platitudes. Of course, we could package in nonnegotiability by including one or more platitude in every disjunct, but we need not. Further, as Kriegel notes, 'we can imagine a Lewis sentence in which some platitudes appear in many more disjuncts than others. This would reflect their greater centrality or 'weight' within the concept' (17).

Kriegel's point in discussing various ways to formulate analyses via Ramsey sentences is to illustrate, and demonstrate, just how flexible serious metaphysics-style conceptual analysis can be.[iv] How does this relate to x-phi? Kriegel claims that 'our concepts are much more complex and flexible than traditional conceptual analaysis has assumed,' (22) with the following interesting result.

> The great majority of concepts, I suspect, are such as to require extraordinarily complex Ramsey sentences to capture. Their capturing is a sort of *labor-intensive conceptual analysis*. My suggestion is that the rationale for experimentalization is the need for this kind of labor-intensive conceptual analysis. Given the complexity of most of our concepts, capturing their psychological structure in full detail would require a multitude of teams of researchers working in parallel to (a) produce hypotheses about aspects of Ramsey sentences, (b) devise the thought experiments that could test the hypotheses, and (c) implement the tests through the familiar social-psychological-style questionnaires presented to the right kinds of subject. (23)

As Kriegel no doubt recognizes, when practicing labor-intensive conceptual analysis, the analyst will come to face difficult decisions. Which platitudes should we regard as central, and which as peripheral? How are we to decide if some concept contains a nonnegotiable element? And so on. Kriegel says little about how the theorist might find guidance with respect to such decisions, but presumably, given the background motivations of serious metaphysics, fidelity to widespread intuitions or conceptual applications across various cases will be important. Given this presumption, evidence of the sort presented above might be taken to support an incompatibilist analysis of free will. On such an analysis, an incompatibility between free will and determinism will take central stage, while considerations closer to the compatibilist's heart will be pushed towards the periphery.

This understanding of the relationship between our concept and the proper concept is in line with that of many philosophers of action, and also with what many take to be a primary motivation for engaging in the experimental philosophy of free will. A different view, however, is available.

In a paper co-written with James Justus, he and I offer the following five contributions x-phi makes to conceptual understanding.

> X-phi can (i) uncover regions of vagueness in extensions and intensions of concepts . . . (ii) reveal instances of conceptual pluralism underlying a notion . . . (iii) discover sources of bias that influence intuitions . . . (iv) discover unpredictable influences on conceptual judgments . . . (v) outline a concept's central features and its dependence relationships with other concepts. (390)

These are all contributions that would, it seems, be equally welcome to the Kriegelian practicing labor-intensive conceptual analysis. And yet we present these contributions as constituting 'a key element in a defensible contemporary alternative to conceptual analysis,' an element they call *explication preparation* (390). The term explication is due to Rudolph Carnap (1950), who argued that theoretical progress in science often requires some measure of conceptual revision, usually in the direction of increased precision or simplicity. Regarding Carnap's motivation, we comment: 'Prioritizing precision and fruitfulness over strict preservation of conceptual content reflects methodology in science and as the unparalleled exemplar of epistemic success in human inquiry Carnap thought philosophy should follow suit' (388).

Justus and I offer Carnap-style pragmatic arguments in favor of the explication of concepts with straightforwardly empirical content, as well as the concepts that concern formal epistemologists. We do not speculate regarding concepts with explicitly moral or otherwise value-laden content, nor do they consider the possibility that some concepts – perhaps free will is an example, due to its intimate connections with moral responsibility – might have content that is both empirical and in some sense normative. Nonetheless, it is hard to deny that scientific and philosophical theory-building bears two features we repeatedly emphasize. First, when theory-building there is a frequent need to move beyond the kind of content implicit in the patterns folk (or even specialist) intuitions indicate. Second, when making such a move some justification must be offered. Carnap, and accordingly Justus and I, emphasize higher-level theoretical justifications, such as fruitfulness in ensuring empirical adequacy and in generating accurate predictions.

Kriegel's proposal on the one hand, and Justus's and mine on the other, differ with respect to the metaphysical pictures that animate them. Kriegel appeals to the

serious metaphysics of the Canberra Plan. We appeal to the deflationary metaphysics of Carnap. Even so, these two proposals appear to share the following features. On either proposal one will need to make difficult choices between various ways to elucidate a concept's potentially quite complex structure. Accordingly, on either proposal one will need guidance, and one will need some justification for the choices made. Here various other features of normal philosophical practice are likely to come into play – arguments that appeal to formal coherence, or to background metaphysical pictures, that give weight to one element of a concept over another, that reinterpret one element of a concept in the name of overall plausibility, or theoretical unity, or reflective equilibrium, or whatever. The two proposals I have considered may place weight at very different places when making such decisions, but they need not. As Justus and I comment,

> [T]he contribution x-phi makes will not determine, in any particular case, how explication should go. Explicative choices (e.g., choices about which features of concepts to preserve and which to abandon) will be guided in part by theoretical aims particular to the case at hand. (391)

We emphasize not only the guidance one's context-sensitive theoretical aims can provide, but also the fact that, for Carnap, fidelity to a concept's intuitive content was far less important than that the conceptual structures produced by explication serve one's aims. This last emphasis raises a distinct possibility. Depending on how the arguments go, it may be fruitful for an analysis to leave behind elements of our concept of free will.

This possibility is explicitly advanced by Manuel Vargas (2013), who argues that no matter our concept of free will, the proper concept is one that leaves our concept behind – that changes its content – for reasons proprietary to moral theory. According to Vargas, our concept of free will contains certain conflicting strands: 'we have diverse intuitions, and some of those intuitions are plausibly understood as compatibilist, and others as incompatibilist' (22). Furthermore, the way forward is not to attempt to capture the complex structure of our concept as accurately as possible, but rather to revise our concept in ways that promote a system of moral responsibility practices that promote morally governed agency.

Now, whether Vargas's revision (we might call it explication) of FREE WILL ought to be accepted depends on a wide range of considerations. We should debate whether the needs of moral theory (or, indeed, of practical life) should guide conceptual revision in the way Vargas indicates. We should debate whether anything is lost in such explicit rejection of our concept. These are debates we can and will have no matter the psychological structure of free will.

This is not to say, however, that empirical work on the psychology of free will be irrelevant to these debates. Indeed, given the complexity inherent in our concept of free will, it looks as though these debates will be difficult to have without much more in the way of empirical work and psychological theorizing regarding the structure of our concept of free will, and much more in the way of effort mapping this structure to the possible – and, one might add, the best – structures the proper concept of free will might take.

Acknowledgements

Thanks to David Rose and an anonymous referee for comments. This work was supported by Wellcome Trust Investigator Award 104347/Z/14/Z.

References

Alexander, J., Mallon, R., and Weinberg, J.M. (2010). Accentuate the negative. Review of Philosophy and Psychology 1(2): 297-314.

Carnap, R. (1950). Logical Foundations of Probability. Chicago: University of Chicago Press.

Kriegel, U. this volume.

May, J. (2015). On the very concept of free will. Synthese 191(12): 2849-2866.

Murray, D. and Nahmias, E. (2014). Explaining away incompatibilist intuitions. Philosophy and Phenomenological Research 88(2): 434-467.

Nichols, S. and Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. Noûs 41(4): 663-85.

Rose, D. and Nichols, S. (2013). The lesson of bypassing. Review of Philosophy and Psychology 4(4): 599-619.

Shepherd, J. (2012). Free Will and Consciousness: Experimental Studies. Consciousness and Cognition 21(2): 915-927.

Shepherd, J. (2015). Consciousness, free will, and moral responsibility: Taking the folk seriously. Philosophical Psychology 28(7): 929-946.

Shepherd, J. and Justus, J. (2015). X-phi and Carnapian explication. Erkenntnis 80(2): 381-402.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. Philosophical Studies 132(1): 99-107.

Vargas, M. (2013). Building Better Beings: A Theory of Moral Responsibility. Oxford: Oxford University Press.

Widerker, D. and Shnall, I. (2015). On the luck objection to libertarianism. In C. Moya, A. Buckareff and S. Rosell (eds.), Agency, Freedom, and Moral Responsibility. Basingstoke: Palgrave-Macmillan.

Appendix

Liberty + / Ensurance + / Consciousness +

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature **needn't** cause the exact same events for the entire history of the universe. So, every time the universe is re-created, **some things may not** happen the exact same way.

Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several psychologists and they assure him that the problem has to do with mechanisms of conscious impulse control in his brain, mechanisms of which he is consciously aware. If Mr. Q's conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives

in Universe 32, the initial conditions of the universe and the laws of nature **do not** determine what he will do. Two, there is the issue of whether Mr. Q's conscious impulse control mechanisms are focused in the right way.

As it happens, Mr. Q's impulse control is consciously focused, and he refrains from uttering the obscenity.

Liberty + / Ensurance + / Consciousness -

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature **needn't** cause the exact same events for the entire history of the universe. So, every time the universe is re-created, **some things may not** happen the exact same way.

Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several psychologists and they assure him that the problem has to do with mechanisms of non-conscious impulse control in his brain, mechanisms of which he has no conscious awareness. If Mr. Q's conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q

utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives in Universe 32, the initial conditions of the universe and the laws of nature **do not** determine what he will do. Two, there is the issue of whether Mr. Q's non-conscious impulse control mechanisms are focused in the right way.

As it happens, Mr. Q's impulse control is non-consciously focused, and he refrains from uttering the obscenity.

Liberty + / Ensurance - / Consciousness +

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature **needn't** cause the exact same events for the entire history of the universe. So, every time the universe is re-created, **some things may not** happen the exact same way.

Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several psychologists and they assure him that the problem has to do with mechanisms of conscious impulse control in his brain, mechanisms of which he is consciously aware. If Mr. Q's conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives in Universe 32, the initial conditions of the universe and the laws of nature **do not** determine what he will do. Two, there is the issue of whether Mr. Q's conscious impulse control mechanisms are focused in the right way.

As it happens, Mr. Q's impulse control is not consciously focused, and he utters the obscenity.

Liberty + / Ensurance - / Consciousness –

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature **needn't** cause the exact same events for the entire history of the universe. So, every time the universe is re-created, **some things may not** happen the exact same way.

Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several psychologists and they assure him that the problem has to do with mechanisms of non-conscious impulse control in his brain, mechanisms of which he has no conscious awareness. If Mr. Q's conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives in Universe 32, the initial conditions of the universe and the laws of nature **do not** determine what he will do. Two, there is the issue of whether Mr. Q's non-conscious impulse control mechanisms are focused in the right way.

As it happens, Mr. Q's impulse control is not non-consciously focused, and he utters the obscenity.

Liberty - / Ensurance + / Consciousness +

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature **will** cause the exact same events for the entire history of the universe. So, every time the universe is re-created, **everything will** happen the exact same way.

Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several psychologists and they assure him that the problem has to do with mechanisms of conscious impulse control in his brain, mechanisms of which he is consciously aware.

If Mr. Q's conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives in Universe 32, the initial conditions of the universe and the laws of nature **determine** what he will do. Two, there is the issue of whether Mr. Q's conscious impulse control mechanisms are focused in the right way.

As it happens, Mr. Q's impulse control is consciously focused, and he refrains from uttering the obscenity.

Liberty - / Ensurance + / Consciousness –

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature **will** cause the exact same events for the entire history of the universe. So, every time the universe is re-created, **everything will** happen the exact same way.

Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several psychologists and they assure him that the problem has to do with mechanisms of

non-conscious impulse control in his brain, mechanisms of which he has no conscious awareness. If Mr. Q's non-conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives in Universe 32, the initial conditions of the universe and the laws of nature **determine** what he will do. Two, there is the issue of whether Mr. Q's non-conscious impulse control mechanisms are focused in the right way.

As it happens, Mr. Q's impulse control is non-consciously focused, and he refrains from uttering the obscenity.

Liberty - / Ensurance - / Consciousness +

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature **will** cause the exact same events for the entire history of the universe. So, every time the universe is re-created, **everything will** happen the exact same way.

Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several

psychologists and they assure him that the problem has to do with mechanisms of conscious impulse control in his brain, mechanisms of which he is consciously aware. If Mr. Q's conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives in Universe 32, the initial conditions of the universe and the laws of nature **determine** what he will do. Two, there is the issue of whether Mr. Q's conscious impulse control mechanisms are focused in the right way.

As it happens, Mr. Q's impulse control is not consciously focused, and he utters the obscenity.

Liberty - / Ensurance - / Consciousness –

Imagine there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same initial conditions and the same laws of nature **will** cause the exact same events for the entire history of the universe. So, every time the universe is re-created, **everything will** happen the exact same way.

Consider a person, Mr. Q, living in this kind of universe (Universe 32). When Mr. Q gets angry, he sometimes utters horrible obscenities. Mr. Q does not like this about himself, and struggles to control it. But he sometimes fails. He has been to several psychologists and they assure him that the problem has to do with mechanisms of non-conscious impulse control in his brain, mechanisms of which he has no conscious awareness. If Mr. Q's conscious impulse control mechanisms are focused in the right way, he has enough control to refrain from uttering horrible obscenities.

Consider one incident in particular. At 10:41 a.m., Mr. Q's boss says something that makes him angry. If Mr. Q utters an obscenity now, he will get fired. Does Mr. Q utter the obscenity or refrain? Two factors might be relevant. One, since Mr. Q lives in Universe 32, the initial conditions of the universe and the laws of nature **determine** what he will do. Two, there is the issue of whether Mr. Q's non-conscious impulse control mechanisms are focused in the right way.

As it happens, Mr. Q's impulse control is not non-consciously focused, and he utters the obscenity.

---

[i] Both proposals have been challenged. For a response to Nichols and Knobe, see Sosa (2007). For a response to Murray and Nahmias, see Rose and Nichols (2013).

[ii] A referee correctly notes that the standard view in cognitive science is that very few concepts have classical structure, and thus that what is of interest here are the particular features that influence applications of the concept free will.

[iii] In connection with this point, one might wonder why, given the presence of a third co-important feature within the psychological structure of our concept of free will (namely, consciousness), there is no corresponding traditional view regarding free will.

[iv] A second interesting alternative on the Ramsey sentence – called a (mega)Lewis sentence – is introduced in response to the idea that some concepts contain a conditional structure. See Kriegel for discussion.