Reactions & Debate

Just War

Jeff McMahan – Rutgers University, USA

A response to Michael NEU: "Why McMahan's *Just* Wars are only *Justified* and Why That Matters." *Ethical Perspectives* 19/2 (2012): 235-355.

Michael Neu is right in claiming that the distinction I have drawn between just and justified wars fails to cohere with another distinction I have drawn between Just Threats and Justified Threats. My failure to see this was a mistake and I am grateful to him for making me aware of it. In this short response to his paper, I indicate how I propose to correct the mistake. My proposal differs from his own recommendation. It is important to be clear and precise about these matters, though I am unconvinced that there is as much at stake as Neu suggests.

Neu points out that I have followed traditional usage in understanding a just war as one that has a just cause and satisfies other relevant conditions of *jus ad bellum*, and that I have referred to those who fight in such a war as 'just combatants'. A just cause for war, as I have understood it, is an aim that it is justifiable to pursue by means of war on the ground that those whom it is necessary to attack as a means of achieving that aim are both morally liable to be attacked for that purpose and sufficiently numerous to make military action necessary. As Neu also observes, I have sought to introduce a further term – 'justified war' – to refer to wars in which some or all of those whom it is necessary to attack as a means of achieving the war's aims are not liable to attack, but for which there is nevertheless a lesser-evil justification – that is, a justification based on the fact that the expected harms or wrongs that the war would avert would be substantially greater than those it would cause. So instead of settling for the simple traditional dichotomy between just and unjust wars, I have suggested that we should distinguish among just, justified, and unjust wars.

Neu contrasts these definitions with those I have given for a couple of other terms I introduced. I have defined a Just Threat as someone whose threatening action will

ETHICAL PERSPECTIVES 19, no. 2(2012): 257-276.

doi: 10.2143/EP.19.2.2160706

harm people only in ways in which they are liable to be harmed, and a Justified Threat as someone whose action, though morally justified, will harm people in ways to which they are not liable. A Justified Threat might be someone, such as the familiar tactical bomber, whose action will intentionally inflict harms to which the victims are liable, but also unintentionally cause other harms to which the victims are not liable. Or a Justified Threat might be someone who has a lesser-evil justification for intentionally inflicting harms to which the victims are not liable.

As Neu rightly comments, these various definitions do not cohere. They use the term 'just' in different ways. Applied to those I have called 'Threats', the term refers exclusively to those whose action will inflict harms only on those who are liable to suffer those harms. Yet wars to which I apply the term 'just' almost invariably involve the infliction of some harms to which the victims are not liable. And this is often true as well of acts of war by those to whom I refer as 'just' combatants. Thus, in distinguishing between different types of Threat, I use 'just' to describe agents and acts that inflict only harms to which the victims are liable, while in discussing wars and combatants, I use it to refer to agents, acts, and courses of action that inflict both harms to which the victims are liable and harms to which the victims are not liable. This is both confusing and confused.

Neu believes that 'just' should be reserved for agents, acts, and courses of action, such as wars, that do not inflict harms to which the victims are not liable. He assumes that if an act harms an individual in a way in which he or she is not liable to be harmed – that is, if it wrongs or infringes the rights of any individual – it cannot be considered just. He therefore thinks that the term 'just war' can coherently be applied only to wars that will not inflict harms to which the victims are not liable – or, perhaps, to wars that can reasonably be expected not to cause such harms. As he phrases it, "a just war, then, according to McMahan's distinction between just and justified threats, is an act which does not wrong people" (240). Because at present and for the foreseeable future such a war is in practice impossible, Neu concludes that there are no just wars. There are at best only justified wars, in which some or all of the harms caused are ones to which the victims are not liable. And because there are no just wars, there are also no just combatants, given the way I have defined that term – though combatants who fight for a just cause may be Just Threats when their action will inflict only harms to which the victims are liable.

I think we should resist Neu's way of understanding these terms mainly because we should, so far as possible, speak the same language as our predecessors in the just war tradition. Writers in this tradition have always known that it is virtually impossible to fight a war without physically harming people who are innocent in the sense of not being morally liable to be physically harmed. Yet despite this knowledge, they have consistently referred to certain wars as just wars. Just war theorists have thus always applied the term 'just war' to wars they knew involved the infliction of unjust harms. Neu claims that in using the term 'just' in this way, they have been guilty of incoherence – that is, of failing to make sense. Yet it seems that we have always understood what they meant; hence their way of using the word 'just' must be intelligible. What is incoherent is my use of 'just' in some cases to include acts that harm non-liable people and in other cases to exclude

such acts. In the remainder of this response, I will draw some distinctions that I hope are accurate and precise and will introduce some terminology that is at least an improvement over that which I have used previously.

It seems that action that causes significant harm, but that can nevertheless be described as just must have a justification grounded in desert or liability. (Because I think that desert has only a small role, if any, in justifying acts of war, I will not discuss desert-based justifications for harming.) War, as I have conceded, causes significant harms that cannot be justified on grounds of liability. Yet, following standard usage in just war theory, I suggest that it is sufficient for a war to be just that there is a liability justification (or, perhaps, that it is reasonable to believe that there is a liability justification) for all the harms that it is necessary to cause as an intended means of achieving the war's aims. A just war, in other words, is one in which the harms that must be intended as a means are justified on the ground that the victims are liable to suffer those harms, while the harms that are expected to be caused to those who are not liable to suffer them are unintended effects that are justified on the ground that they are the lesser evil. This is the way the term 'just war' has always been used within the just war tradition. What is new, perhaps, is the explicit recognition that all just wars require two distinct forms of justification: a liability justification for the intended harms and a lesser-evil justification for the unintended harms caused to those who are not liable to suffer them.

It may be helpful to mention briefly the way in which this characterization of a just war is related to the notion of a just cause for war as I understand it. For simplicity of exposition, I will consider only a just cause for *defensive* war, ignoring the possibility that the rectification of a wrong that has already been committed could in some cases be a just cause for war. For there to be a just cause for defensive war, there must be a threatened wrong for which many people would be responsible, either directly or indirectly through their commitment to assist or to protect the primary wrongdoers if necessary. It must also be the case that the harms that it would be necessary to inflict as a means of preventing the wrong are ones to which the victims would be morally liable by virtue of the nature of the wrong and the degree of their responsibility for it. This means that the threatened wrong must be sufficiently significant to make those who are or would be responsible for it potentially liable to be killed as a means of preventing it. Finally, for the just cause to be a just cause *for war*, it must be necessary to attack sufficiently many people that only military means could be effective in achieving the justifying aims. This account of a just cause for war explains the way in which the insistence that a just cause is necessary for a just war guarantees that that harms that must be intended as a means of achieving the aims of a just war must have a liability justification.

Neu of course has a point: it could be potentially misleading to apply the label 'just war' to wars that involve harms that are unjust – that is, harms that wrong or infringe the rights of the victim. We could avoid this problem by introducing some finer-grained distinctions. We might, for example, divide what are standardly referred to as just wars into two distinct categories. Wars in which no one is harmed unjustly could be called *perfectly just wars*, while those in which there was a liability justification for the necessary

intended harms, a lesser-evil justification for all unintended harms to non-liable people, and no unnecessary harms could be called *imperfectly just wars*. But this seems unhelpful, in part because there are hardly any occasions for the use of the term 'perfectly just war' and in part because 'imperfectly just war' is clumsy and unnecessary given that it is possible to guard against misconceptions by explicitly noting that the term 'just war' applies, as it always has, to wars that are imperfectly just in the sense in that they foreseeably involve the infliction of some unintended but nevertheless unjust harms.

The question now arises whether we should make parallel claims about cases of individual action. We might, for example, define Just Threats as those whose action intentionally causes harm only to those who are liable but may also cause harms to which the victims are not liable, but for which there is a lesser-evil justification. According to this definition, the tactical bomber is a Just Threat.

An alternative possibility would be to relativize a threatening person's status to the status of his or her potential victims. The tactical bomber, for example, might be a Just Threat vis-à-vis the intended victims of his or her action, yet also be a merely Justified Threat vis-à-vis the innocent civilians his or her action would unavoidably harm as an unintended but proportionate side effect.

Neither of these suggestions is ideal. The first obscures significant differences among actual cases and the second is unnecessarily awkward. In cases of individual action, clarity is best achieved by distinguishing four categories, the first two of which correspond to the two types of war that I said earlier we should combine under one label: namely, perfectly just wars and imperfectly just wars. The four categories are as follows.

i. A *perfectly just threatener* is one who would cause harms, whether intentionally or unintentionally, only to people who are liable to suffer those harms.
ii. An *imperfectly just threatener* is one who would intentionally harm only those who are liable to be harmed but would also unavoidably cause unintended but proportionate harm to some people who are not liable.
Imperfectly just threateners might, for brevity, be referred to as *just threateners*.
iii. A *justified threatener* is one who has a lesser-evil justification for intentionally harming some people who are not liable to be harmed, or for harming some people who are liable to be harmed by more than the amount of harm to which they are liable.
iv. An *unjust threatener* is one who would inflict harms without moral justification.

Even though the term 'perfectly just threatener' is cumbersome, the main reasons for not adopting the parallel term 'perfectly just war' do not apply. First, while in the foreseeable future there are unlikely to be any instances of a war in which no one is unjustly harmed, there are many instances in which an individual can inflict a harm to which the victim is liable without harming anyone else as a side effect. Second, there is no alternative label that is already widely used that we ought to try to preserve – no parallel, in other words, with the label 'just war'."

Just combatants may at different times fit into any of these categories. They are perfectly just threateners if their act of war would harm only unjust combatants, just threateners if their action would harm unjust combatants but also cause unintended but proportionate harm to innocent bystanders, justified threateners if they have a lesser-evil justification for intentionally harming innocent bystanders, or unjust threateners if they harm people without justification. If just combatants become unjust threateners, they may be war criminals or murderers.

I will conclude by registering agreement with much of what Neu says in the final section of his critique. Indeed, I would go even farther than he does. Much of the material in his final section is premised on his earlier claim that people "are *not* generally required to prevent situations in which they can act justly" (237). That may be true in general but it is not true of fighting a just war. For the sake of illustration, consider the parallel with individual self-defence. Suppose that if I remain in my home, I will be confronted by a murderous aggressor whom I will then have a liability justification for killing in self-defence, for at that point my options will be limited to two: killing the aggressor of being unjustly killed by the aggressor. Assuming that if I flee from my home I will be able to ensure that the aggressor will later be captured by the police, I am morally required to retreat from the confrontation, thereby preventing the situation in which I would be a perfectly just threatener and could act justly in killing the aggressor. Some people maintain that I have a right to remain in my home in all such cases, but that is a mistake. I may remain in my home only if my being in my home at that particular time is so important that killing a culpable intruder would be a proportionate means of enabling myself to stay there.

It therefore seems that, contrary to Neu's assumption, there is normally a strong moral reason to avoid conditions in which it would be permissible or even morally required to fight a just war – *even* a just war that would harm people only in ways to which they would be liable (this may even be an implication of the *ad bellum* requirement of necessity). If this is correct, it weakens Neu's reasons for concern about the way that I and others use the term 'just war'.

We Can Test the Experience Machine

Dan Weijers – Victoria University of Wellington, New Zealand

A response to Basil SMITH: "Can We Test the Experience Machine?" *Ethical Perspectives* 18/1 (2011): 29-51.

In his provocative "Can We Test the Experience Machine?", Basil Smith argues that we should recognise a limit on experimental philosophy. In this response to Smith, I will argue that his limit does not prevent us from usefully testing most experience machine thought experiments, including De Brigard's inverted experience machine scenarios.

I will also argue that, if taken seriously, Smith's limit has far-reaching consequences for traditional (non-experimental) philosophy as well.

Smith describes his proposed limit on experimental philosophy as follows:

> *[C]ertain* philosophical thought experiments (e.g. the experience machine, the trolley problem, etc.) ask subjects to make decisions from the position of *confronted agents*, or those who have entered a specific state of mind. But when taking surveys (answering either 'yes' or 'no' or giving a score on a Likert scale), subjects are *not* in that state of mind, nor can they imagine it. Therefore, when experimental philosophers claim they have tested certain philosophical experiments (and thereby the intuitions they evoke) we have reason to believe they have not (30; italics original).

Essentially, Smith recommends that experimental philosophers should stop asking questions of the form: 'how *would* you react in this situation?' (henceforth *would* questions), especially when the situation is hypothetical, intense, and unfamiliar. I will refer to this as 'Smith's limit'. Smith argues for his limit on the grounds that survey answers to *would* questions do not provide useful information because they require respondents to perform a very difficult task – accurately predicting their reactions to hypothetical, intense, and unfamiliar situations.

Smith gives two reasons for why it is so difficult for survey respondents to accurately predict their reactions to hypothetical, intense, and unfamiliar thought experiments. Both of the reasons are related to how difficult it is for survey respondents to adopt the role of an appropriately confronted agent. A confronted agent is someone who has context-specific subjective experiences, including affective responses such as feelings of confusion, incredulity, fear, and uncertainty. First, the respondents would have none of the affective responses they would have experienced if the scenario were real (39). Second, respondents would have given their ideal responses (how they think they *should* react in the given scenario) as opposed to how they think they would actually react if they really found themselves in the hypothesised scenario (39). Since the affective responses we experience can heavily influence our reactions, and since how we think we should react is often different to how we would actually react, Smith concludes that "we *cannot compare the responses subjects give on a survey with their reactions to the real event*" (39; italics original).[1]

Smith also provides two examples of types of studies that require respondents to adopt the role of confronted agents *to some extent*; studies that involve respondents anticipating their own futures and ones that involve respondents attempting to identify with the moral decisions of others (46). Smith concludes his article by clarifying that his limit on experimental philosophy affects studies "*just to the degree*" that they require respondents to adopt the role of confronted agents (46; italics original). Presumably this means that studies with questions requiring respondents to undertake just a small amount of *would*-based pondering (relative to the total pondering required to form a judgment in about relevant scenario) should not automatically be considered failed tests, just somewhat undermined tests.

Smith acknowledges that his limit "does not apply to most experimental studies" (46). But does it apply to studies on experience machine thought experiments? Or, as Smith puts it in the title of his paper: can we test the experience machine? Although never directly stated, we can infer that Smith's answer to this question would be: 'we can only test experience machine thought experiments that do not require respondents to adopt the role of a confronted agent'. Smith is very clear about his belief that De Brigard's inverted experience machine thought experiments cannot be tested:

> [T]he inverted experience machine, as well as other similar such experiments (e.g. justified theft dilemmas, questions of torture, the trolley problem, etc.) have a unique set of characteristics that make it impossible to gather the right subjects to test [i.e. subjects that can adopt the role of an appropriately confronted agent]. Therefore, in practice, these thought experiments are impossible to test (37).

Smith is relatively quiet, however, on whether other variants of the experience machine thought experiment can be tested. I will argue that his limit does not prevent us from usefully testing most experience machine thought experiments, including De Brigard's inverted experience machine scenarios.

The most famous experience machine thought experiment was proposed by Robert Nozick. Nozick describes experience machines as being able to "give you any experience you desired […] a lifetime of bliss" all without you realising that your experiences were machine-generated (1974, 42-43). Nozick then asks his readers: "Would you plug in" to an experience machine for the rest of your life? (1974, 43). This is clearly a *would* question – one that, in Smith's view, only an appropriately confronted agent could credibly provide an answer to. Therefore, Smith would object to any attempt to test this scenario.

De Brigard asks "What would you choose?" at the end of all of his scenarios (2010, 47). Therefore, applying Smith's limit to De Brigard's inverted experience machine thought experiments reveals the same judgment; Smith would object to any attempt to test this scenario too.

Smith's claim that De Brigard's tests are useless, and his implication that any test of Nozick's experience machine scenario would also be useless, is based on the assumption that these tests have the purpose of discovering what people would choose if they were really offered a choice between reality and a life in an experience machine. It is far from clear, however, that when De Brigard and Nozick were creating their thought experiments, discovering the truth about what people would do was their ultimate aim.

The stated purpose of De Brigard's tests was to investigate how informing participants that they "have been living inside an experience machine [their] entire life […] would affect, per se, their judgements on their own happiness or well-being" through studying their responses to his scenarios (2010, 46). It is true that De Brigard mentions the connection between the experience machine and psychological hedonism: "Many philosophers […] consider the alleged effectiveness of this thought-experiment a rather convincing proof against the tenability of psychological hedonism" (2010, 45).[2] However, this is the only mention of psychological hedonism in the whole article. De Brigard

makes it clearer in his conclusion (2010, 53-55) that his primary research question is whether our preferences in experience machine scenarios are best explained by our valuing of reality or the status quo.

The stated purpose of Nozick's experience machine scenario is to demonstrate that "something matters to us in addition to experience" (1974, 44). Furthermore, when Nozick discusses his experience machine thought experiment in more detail, he stresses that he takes the experience machine to provide evidence about what has prudential value, not just what motivates us: "the connection to actuality is important whether or not we desire it – that is *why* we desire it – and the experience machine is inadequate because it doesn't give us *that*" (1989, 106-107; italics original).

For both of these stated purposes, then, it seems that knowing what people think is in their best interest to do, would be more useful than knowing what they would actually choose. This is because (as Smith has pointed out) when actually choosing, people are confronted agents and their decisions are affected by feelings of confusion, incredulity, fear, and uncertainty. In most cases, we should expect these influences to cause confronted agents' revealed choices to be less rational than they would have otherwise been.[3] However, if we could really know what people think they *should* choose, then we would have a much better understanding of our well-being-related judgements, which is what it seems De Brigard and Nozick were really after.

In fact, assuming that Smith is correct about the reasons why survey participants cannot credibly respond to *would* questions, then De Brigard and Nozick's 'what would you choose?' questions are probably eliciting responses more closely aligned with what the respondents think they should choose if they were given that option. If the feelings of confusion, incredulity, fear, and uncertainty that are required to properly adopt the role of a confronted agent for De Brigard and Nozick's scenarios are likely to cause less rational choices, then the removal of them is likely to lead to more rational choices (i.e. choices that better reflect what the respondents think they should choose).[4] Therefore, while Smith (39) bemoans the fact that responses to De Brigard's scenarios are likely to be the participants "ideal choices" (as opposed to what they would actually choose if the scenario were real), it actually seems like a strength of De Brigard's experiments.

Indeed, using questions of the 'how *would* you react in this situation?' variety is commonly used even when the researchers are not directly interested in what respondents would actually do if the scenarios were real. For example, Petrinovich and colleagues ran experiments on typical philosophical thought experiments (including trolley problems) using explicit *would* questions, but made it clear that they were most interested in better understanding people's underlying moral intuitions (Petrinovich, O'Neill and Jorgensen 1993, 476).[5] Furthermore, Alex Barber has argued that "I read [Nozick's experience machine objection to hedonism] as being interested in your views about what you (self-interestedly) *should* do, but what you *would* do is our best guide to these views" (2011, 263, n. 6; italics original).[6] 'Best guide' is surely too strong here, but this quote shows that philosophers are also aware of the importance of the distinction between what we *would* do and what we *should* do in regards to the experience machine.

So what does this mean for experimental studies of experience machine scenarios that directly ask *would* questions, but do not seek to discover how people would actually react if the scenarios were real (a category that seems to include De Brigard's experiments)? Does Smith's limit deem them useless?

Even if the questions were reworded to (or reinterpreted by non-confronted respondents as) 'what *should* you choose?' questions, De Brigard and Nozick's scenarios still require respondents to contemplate some *would* questions. On the new wordings (or interpretations), respondents no longer have to predict if they would actually choose an experience machine life, but they still have to ask themselves what a life in an experience machine *would* be like. Imagining what an experience machine life would be like is fairly easy for someone who has previously been confronted with one. Unfortunately, no one has come across an experience machine yet, so there is no group of people who could answer the question 'what should you choose?' as an appropriately confronted agent. This is one of the main points that Smith was trying to make (captured in the following claim):

> [T]he inverted experience machine… [has] a unique set of characteristics that make it impossible to gather the right subjects to test. Therefore, in practice, [it is] impossible to test (37).

Nevertheless, it seems like participants' responses to the question 'what should you choose?' require less knowledge and emotional upheaval to be credible than their responses to the question 'what *would* you choose?' So, Smith might conclude that any studies on reworded versions of Nozick or De Brigard's experience machine thought experiments should be considered as flawed, but still useful tests, as opposed to being worthless non-tests.

And what of De Brigard's actual studies and Nozick's actual scenario? Given their stated purposes, and the likelihood of their 'what would you choose?' questions being reinterpreted as 'what should you choose?' questions, then it seems that Smith's limit does not rule out their usefulness. Indeed, it seems like only a very uncommon kind of experimental study on an experience machine thought experiment would be ruled out by Smith's limit. Furthermore, the same arguments that I have run here could be used to redeem most of the other experimental studies that Smith claimed to be failed tests (trolley problems etc.).

All of the results of my application of Smith's limit differ to his application of it. I have argued that appropriately applying Smith's limit identifies a few experimental studies as worthless and many as flawed (but still useful). Smith may disagree with my analysis of which particular studies belong in each of these categories, but we seem to agree on his justifications for being concerned about *would* questions. Non-confronted agents cannot imagine having the intense emotions that *really* being in extreme and unfamiliar situations elicit. And since intense emotions can have strong and unpredictable affects on our reactions, non-confronted agents should not be expected to be able to accurately predict how they would react in extreme and unfamiliar situations.

This is why responses to experience machine scenarios should not be taken at face value – no one could be sure about how they would react to being in an experience machine. So, Smith and I agree that questions requiring survey respondents to predict how they would react in future or hypothetical scenarios are undermined, but perhaps we disagree about the extent to which asking *would* questions affects an experiment's usefulness.

When considering the extent to which asking *would* questions will affect a study's usefulness, it is worth also thinking about a very closely analogous case. Notice that when philosophers, and other readers of philosophy, ponder thought experiments, they too are required to predict how they would react in future or hypothetical scenarios or attempt to identify with the moral decisions of others. Notice also that philosophers and philosophy students do not usually appear to be experiencing feelings of confusion, incredulity, fear, and uncertainty when they read or discuss Nozick's experience machine or any other thought experiment.[7] Smith makes no comment about his limit on experimental philosophy applying to philosophy in general, but his justifications for the limit seem to apply to nearly all of philosophy. Many arguments in philosophy use thought experiments that require the reader to either predict how they would react in future or hypothetical scenarios or attempt to identify with the moral decisions of others. Therefore many arguments in philosophy would be deemed either flawed or useless by Smith's limit.

I repeat Smith's justification of his limit here, with a few words changed, to show how easily it applies to much more than just experimental philosophy:

> *[C]ertain* philosophical thought experiments (e.g. the experience machine, the trolley problem, etc.) ask subjects to make decisions from the position of *confronted agents*, or those who have entered a specific state of mind. But when [reading thought experiments, readers] are *not* in that state of mind, nor can they imagine it. Therefore, when [people] claim they have [learnt from] certain philosophical [thought] experiments (and thereby the intuitions they evoke) we have reason to believe they have not (30; italics original).

Of course, there are other reasons why the results of experimental studies may be less reliable than the opinions of philosophers about a particular thought experiment.[8] But Smith discusses his limit on experimental philosophy separately, claiming that it is strong enough *by itself* to show that certain thought experiments cannot be tested at all (37). Given that the justifications for Smith's limit imply that all thought experiments are undermined to the extent that they require us to adopt the role of confronted agents, I am curious about what Smith believes. He might believe: that philosophers – and others who read thought experiments – are better at adopting the role of confronted agents, or that large swathes of philosophy are flawed or useless, or that Nozick and De Brigard's scenarios (whether experimented on or simply read by someone) are only slightly undermined by his concern about adopting the role of an appropriately confronted agent.

Works Cited

Barber, Alex. 2011. "Hedonism and the Experience Machine." *Philosophical Papers* 40/2: 257-278.

De Brigard, Felipe. 2010. "If You Like it, Does it Matter if it's Real?" *Philosophical Psychology* 23/1: 43-57.

Hodgson, Geoffrey M. 1988. *Economics and Institutions*. Cambridge: Polity.

Nozick, Robert. 1974/1991. *Anarchy, State, and Utopia*, Oxford: Blackwell.

Nozick, Robert. 1989. *The Examined Life: Philosophical Meditations*. New York: Simon & Schuster.

Petrinovich, Lewis, Patricia O'Neill and Matthew Jorgensen. 1993. "An Empirical Study of Moral Intuitions: Toward an Evolutionary Ethics." *Journal of Personality and Social Psychology* 64/3: 467-478.

Smith, Basil. 2011. "Can We Test the Experience Machine?" *Ethical Perspectives* 18/1: 29-51.

Weijers, Dan. 2011. "Hedonism." *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/hedonism/ [accessed March 30, 2012].

Notes

1. Given the justifications Smith provides, this conclusion is too strong. Just how useless the inclusion of *would* questions will make a study seems to be a question of degree even when the *would* questions are explicit and involve participants attempting to predict their reactions to a hypothetical, intense, and unfamiliar situations. If a group of people were asked if they would accept an offer of $1million to attempt a dangerous high-altitude balancing stunt, they would probably not exhibit any signs of feeling confused, incredulous, fearful, or uncertain, but their answers would likely be fairly predictive of their actual choice if the offer were real. Therefore, even though this group would be being asked to predict their reactions to a hypothetical, intense, and unfamiliar situation, and they failed to adopt the role of an appropriately confronted agent, their responses would be far from useless. My goal in this response is to examine the implications of Smith's limit, rather than to question it, however, so this issue will be overlooked here.

2. Psychological hedonism (also known as motivational hedonism) is the theory that all human behaviour is guided by the desires to encounter pleasure and to avoid pain (Weijers 2011).

3. See, for example, Hodgson (1988, 10; 11; 57) on the many revealed 'decisions' based on unconscious reactions to context and psychological states of stress, fear, excitement, social pressure, and so on.

4. I am not making the false claim that rational decisions are those made with as little emotional processing as possible, just that decisions made under normal emotional conditions are likely to be more rational than those made under extreme emotional conditions.

5. "It also might be objected that this research concerns only intuitions and that it is not possible to determine whether the intuitions revealed here would be translated into action. Of

course, this objection is justified. However, the intent of this research is not to understand or predict what people would do but to probe the structure of their systems of moral intuitions. The impetus for the research was provided by the views of many moral philosophers regarding how people construct the world of morality. We argue that these dilemmas probe the core of moral beliefs. It would seem that an understanding of how people resolve these fantasy dilemmas might be a good basis on which to begin to understand action but that translation is not part of the present undertaking, although it would be a fascinating question for future research" (Petrinovich, O'Neill and Jorgensen 1993, 476).

6. Barber goes on to congratulate Nozick for revealing to hedonists that their theoretical convictions might contradict their professed hypothetical actions since it might be able to get hedonists to admit that in their "heart of hearts" they are not really hedonists (2011, 266). Barber summarises: "Our hypothetical actions speak louder than our theoretical words" (2011, 266).

7. I think that the reality here is that imagining a thought experiment does cause affective responses in the imaginer, but that these responses are much more muted than they would have been if the scenario was really experienced.

8. Smith (2011) discusses many potential problems with experimental studies, all of which can be minimised or avoided by sound experimental design.

Affect, Rationality, and the Experience Machine

Basil Smith – Saddleback College, Mission Viejo, USA

A response to Dan WEIJERS: "We Can Test the Experience Machine." *Ethical Perspectives* 19/2 (2012): 261-268.

I. INTRODUCTION

In "Can We Test the Experience Machine?" I address recent tests (i.e. surveys) of the 'inverted' experience machine, conducted by Filipe De Brigard. In this scenario, subjects were supposed to imagine they have been living in an experience machine *for their entire lives* yet it has all been a mistake (De Brigard 2011, 46). I argue that to test how subjects would react to this, such subjects would have to be in the proper affective state (i.e. experience confusion, incredulity, fear, and uncertainty). De Brigard, therefore, tested the wrong subjects. Moreover, I argue that, generally, when subjects with the wrong affective state respond to surveys, they do so as their *ideal* selves – as how they *want* to see themselves. De Brigard did not test control for this, and again tested the wrong subjects (Smith 2011, 45). Since subjects considering the experience machine cannot adhere to these two limits (i.e. being in the right affective state *and* not answering as their ideal selves), the machine cannot be tested.[1]

Dan Weijers, in his "We Can Test the Experience Machine," concedes that, if subjects were to imagine the experience machine, they would have "feelings of confusion, incredulity, fear, and uncertainty" (262). Indeed this would have "strong and unpredictable effects" on their estimations of their futures (265). Even so, Weijers objects to my two limitations (i.e. concerning affect and ideal selves). Without such limitations, he says, subjects would make "more rational choices," and presumably be more reliable.[2] Therefore, my argument is subject to a dilemma: either these two limitations "do *not* prevent us from usefully testing" the experience machine and similar matters, or if taken seriously, they "apply to nearly all of philosophy" (Weijers 2012, 266).

Concerning the first horn of this dilemma, Weijers incurs three problems. First, Weijers, it seems, construes Nozick and De Brigard as having purposes they do not have, such that their actual purposes – they are clear about these – make such testing either irrelevant, or subject to limits. Second, by insisting that experience machine subjects be rational, Weijers *assumes* his conclusion, and also, if such strictures were taken seriously, they would render it impossible to test many subjects we care to. Third, Weijers insists that such experience machine studies illuminate our "judgments about our own happiness" (263), and so are useful, but does not say *how* this is. Since this is so, he only promises an argument, and does not offer one.

Concerning the second horn of this dilemma, Weijers is correct that my two limits "apply to nearly all of philosophy" (266), but he misses the significance of this. First, many thought experiments (e.g. Mary the colour scientist from Frank Jackson, or zombies from David Chalmers), do not reflect on us, our values, or characters. Philosophers offer such thought experiments, moreover, not to test them, but to elicit our intuitions, to refute a theory, etc. Therefore, while my two limits apply widely (e.g. to thought experiments about morality), they do not affect all of philosophy. However, as we will see, these two limits apply to philosophical thought experiments about morality (e.g. about theft, torture, or our commitments to our fellows). Therefore, many such thought experiments are impossible to test.

II. Purpose, Rationality and Use?

Weijers begins by saying that, when understood properly, the original experience machine thought experiment from Robert Nozick, as well as the inverted experience machine studies, can escape my limitations. The claim that these studies are useless is

> […] based on the assumption that these tests have the purposes of discovering what people *would* choose if they were really offered a choice between reality and a life in the experience machine (263).

However, Weijers says it is "[…] far from clear," that "discovering the truth about what people would do" is their ultimate aim (263). In point of fact, De Brigard (and most

experimental philosophers) ask 'would questions' of subjects to discover their *robust* intuitions, or to reveal the *stable* features of their character that explain *why* they answer the way they do, and so explain their behaviour (Kauppinen 2007, 110).[3]

Fortunately, Nozick and De Brigard state their purposes clearly. Against psychological hedonism (i.e. the theory that all that matters to us is pleasure), Nozick says "we want to *do* certain things, and not just have the experience of doing them." Moreover, we "[…] want to *be* a certain way, to be a certain kind of *person*," and there no answer in the machine about whom we are. Lastly, Nozick says the experience machine "[…] limits us to a reality no deeper or more important than that we can construct" (1974, 43). Nozick, then, argues that contact with reality is the *condition of possibility* of our being actors, persons, and our having deep, meaningful lives. Therefore, our choice of the experience machine would be "[…] a kind of suicide" (Nozick 1974, 43; Griffin 1986, 9).

By contrast, De Brigard insists that Nozick and his critics assume that "[…] prefer not plugging in because they value being in touch with reality." However, as he argues, this is not so, since "people are averse to abandon the life they have been experiencing so far, *regardless of whether such life is virtual or real*" (De Brigard 2011, 44). Psychologically, "people do not want to abandon the life they have lived so far, the life they are familiar and comfortable with," regardless (De Brigard 2011, 52). Therefore, when De Brigard asks subjects whether they *would* prefer the experience machine or reality, he hopes to reveal *why*.

Regardless, Weijers says that given their "stated purposes," Nozick and De Brigard can ignore my first limitation (i.e. concerning proper affect). Since subjects who imagine either version of the experience machine would experience "confusion, incredulity, fear, and uncertainty" (262),

> […] knowing what people *think is in their best interest to do* would actually be *more* useful than knowing what they would actually choose (264; italics mine).

Concerning my second limitation, that subjects not answer as their ideal selves, Weijers says "if we could really know what people *think they should choose* then we could have a *much better understanding*" of their happiness judgments (264; iltalics mine).

Weijers says we can rephrase both the original experience machine thought experiment and inverted experience machine studies to ask subjects to answer as their ideal selves, as "[…] how they think they *should* choose."[4] In such a situation "respondents no longer have to predict if they would actually choose the experience machine life" (265). Leaving aside any other difficulties, asking subjects about their ideal selves, about how they should answer, will still yield results. Weijers concludes that, given this potential rephrasing, we can "[…] redeem most experimental studies," such that it is only a "[…] uncommon kind of experimental study" that is truly useless (265).

III. The Importance of Affect

Does Weijers show that the experience machine *can* be tested after all? Unfortunately, he does not, for three reasons. First, Weijers saddles Nozick and De Brigard with purposes they do not have, which are either irrelevant or are still subject to the two limits, cited above. Second, Weijers says that when subjects feel confusion, incredulity, fear, etc., we should test rational subjects instead, but, *beyond his say so*, he gives us no further reasons to infer that their answers would reveal their robust intuitions. Moreover, if psychology were beholden to his strictures about rationality, testing many subjects would be impossible. Third, Weijers insists that experience machine tests illuminate judgments on our own happiness, they are *useful*. But he does not articulate how this is. In other words, he only promises an argument, and so offers no objection. Therefore, Weijers does not show that the experience machine can be tested after all.

Concerning the purpose of the original experience machine thought experiment, Weijers says that Nozick hopes to show that "something matters in addition to experience," as though our responses are crucial. In point of fact, Nozick wants to show that, in addition to experience, we want to be actors, persons, and to have meaningful lives. According to him, only by being in contact with reality can we do, be, or have meaning. This is why Nozick calls the machine "a kind of suicide" (1974, 43). However, this reality condition for doing, being, and meaning *remains in force regardless of how we happen to respond*, even if we choose suicide. Perhaps our responses to the experience machine show that we are not truly hedonists, but such responses are not relevant.

Weijers, oddly, saddles De Brigard with a purpose he clearly does not have. De Brigard asks whether, when subjects are informed that they have been living in the inverted experience machine (i.e. where all their relations, loves, etc., have been "…nothing but the products of a computer program"), this information would make them think their "…lives have *less* value," personally. De Brigard responds that he is "…not even sure whether this piece of information would affect, per se, judgments on their own happiness" (2010, 46). In fact, regardless of what subjects would judge about themselves, De Brigard hopes to show that subjects, although they may appreciate pleasure or reality, "[…] tend to prefer the state they are currently in" come what may. In so many cases, when subjects do not want to connect or to disconnect, this is just "[…] a manifestation of this underlying psychological phenomenon" (De Brigard 2010, 50). De Brigard, then, is not primarily concerned with our "judgments on our own happiness," but hopes to reveal what we truly prefer, despite ourselves.

Unfortunately, this is the problem. How can such testing of subjects in the wrong affective state (i.e. students taking a survey as opposed to being confronted, and so feeling confusion, incredulity, fear, etc., and answering as their ideal selves or as they think they should) ever show if they prefer contact with reality, the pleasure of the machine, or if both of these options are trumped by their wanting to maintain the life they have, regardless? In other words, how can asking such subjects about these options ever show anything about their robust intuitions? Perhaps, as Jason Kawall says, "[…] given

relevantly similar experiences and general knowledge" subjects might predict their own futures (1999, 383). However, in the case of the inverted experience machine, subjects have no such experiences. De Brigard disregards these worries and infers that his subjects prefer maintaining their present state regardless. Clearly, though, the evidence does not show this.

Weijers, on the other hand, postulates that subjects without confusion, incredulity, fear, etc., and answering only ideally or as they think they should, would make "more rational choices." Perhaps so, but how does this help? Does this mean that subjects with little affect, answering only ideally or as they should, are more reliable in some way? Do such rational, so reliable, subjects reveal their robust intuitions. Weijers does not attempt to improve the experience machine studies (e.g. conduct new studies with *more* rational subjects), and so his rational subjects are the same subjects as those of the old studies. Yet this is the problem, for beyond his mere insistence, we have no reason to infer that, since subjects are rational or reliable (i.e. having little affect, and answering ideally or as they think they should), their responses reveal whether they are motivated by contact with reality, the pleasure of the machine, or by the desire to maintain whatever state they are already in, come what may. In other words, we have no reason to take the answers of such as evidence of their robust intuitions about the experience machine. Therefore, when Weijers says that rationality and so reliability mirror said robust intuitions, he assumes his conclusion.

However, could Weijers be correct about rationality, generally speaking?[5] According to his strictures, when subjects are confused, incredulous, fearful, etc., we should exclude them from consideration, and test rational subjects instead. However, applying rationality strictures to typical psychological tests, such as tests of helping behaviour, individual or group responsibility, authority, as well as philosophical thought experiments about morality, what would happen exactly? Stanley Milgram, in an article on the ethics of testing, points out the problem.

> In this situation, only experiments that aroused neutral or positive emotions would be considered ethical topics for experimental investigation. Clearly such a stricture would lead to a very lopsided psychology, one that caricatured rather than accurately reflected human experience (Milgram 1977, 19).

Weijers, in suggesting that subjects with "confusion, incredulity, fear, etc." should be replaced with rational ones instead, is making a *general* claim. Many psychological experiments, thus, as well as many philosophical thought experiments, would be run with rational subjects instead, and would get entirely different results. Clearly, though, when subjects answer here – with the wrong affect, answering only ideally or as they think they should – these results would *not* reveal their robust intuitions about helping, responsibility, authority, or about the thought experiments. It is implausible to suggest that we could get better, or more accurate, results in this way, or even that there is any use to such experiments. Weijers, then, implies that his strictures of rationality are "an advantage,"

but they are actually a hindrance that promises to render the testing of certain subjects arduous, if not impossible.

Lastly, Weijers admits that subjects who imagine the experience machine may have "confusion, incredulity, fear" etc., and that this may have "strong and unpredictable" effects on their choices. Since we can exchange experience machine thought experiment and inverted study 'would questions' for those about our ideal selves or with 'should questions,' the tests are still *useful*. Yet, as we have seen, since this use is clearly *not* that subjects will reveal their robust intuitions about either version of the experience machine, it must be something else. Weijers hints that perhaps such tests are useful because subjects reveal something about "judgments on our own happiness," but says no more. But to claim that the experience machine is useful in this way, and then to not delineate *how* is not offering an argument. Weijers, then, manages to promise an argument, but not to give one.

In conclusion, experimental philosophers say they use "methods of experimental psychology" to test philosophical thought experiments, and so study our robust intuitions (Nadelhoffer and Nahmias 2007, 123; Alexander and Weinberg 2007, 56). Weijers, though, attempts to evade two limitations on such testing (i.e. that subjects be tested in the right affect, and that they not answer ideally or as they think they should) by suggesting that we can test rational subjects instead. Yet, as we have seen, this suggestion offers no changes, and assumes what it has to prove. Moreover, if taken seriously, such strictures on rationality would render the testing of many subjects impossible, past and future. Lastly, Weijers says testing of rational subjects is useful, but never articulates what this use is. Therefore, the experience machine is untestable, in practice.

IV. The Limits of Philosophical Reflection

Concerning the second horn of his dilemma, Weijers says the two limits (i.e. concerning affect and ideal selves, or should answers) are general. Philosophers, though, typically use thought experiments (e.g. Mary the scientist, zombies, etc.), but they and their audience "[…] usually do not *appear* to have feelings of confusion, incredulity, fear and uncertainty." Do we all have the power to overcome these two limits or are our efforts "flawed and worthless" as well?

In point of fact, these two limits do "[…] apply to nearly all of philosophy," but it does not follow that philosophical reflection is worthless. Some thought experiments (e.g. Mary, zombies, etc.) do *not* require us to manifest any specific affect to understand them, but only require them to *imagine* something (e.g. that a scientist knows every physical fact, or that, of two identical brains, one has consciousness and the other does not), and draw a conclusion. Of course, we may fail to imagine what we say we do, or may draw the wrong inferences from whatever we do imagine, but these are different problems (Sorensen 1992, 33-35). Philosophers, by offering these thought experiments, do not hope to test anything, but only to elicit our intuitions, persuade us, or refute theories.

Indeed, such experiments often do get at our robust intuitions about, say, physicalism. Since we are *not* required to have any affects to understand them and since they are *not* testing anything by offering them, the two limits do not apply. Therefore, these philosophical thought experiments and our use of them are not flawed and worthless, after all.

Weijers, though, is correct that these two limits apply to other philosophical thought experiments (e.g. cases about theft, murder, or even communal action) yet seems to miss the significance of this. Philosophers offer these experiments to investigate our values, our character, and our future actions. To really understand, say, the trolley problem, we must either have the experiences of being confronted with the trolley ourselves, or be able to imagine similar experiences and apply them to this case. Indeed, we must have the proper affect, whatever that is. Weijers implies that we "[…] do not appear" to have the proper affect, yet seem to understand. Unfortunately, this merely illustrates how these two limits (i.e. that subjects have the proper affect and not answer ideally or as they should) apply, yet can be easily be ignored. In these philosophical thought experiments, when we have not experienced the situation and cannot conjure the proper affect in any other way, this appearance of understanding is just that. Plainly, we can understand the words when the thought experiment is spoken, but that does not mean we understand it, nor our robust intuitions about it.

V. Conclusion

In conclusion, Weijers argues that either my two limits (i.e. that the subjects of certain philosophical thought experiments be tested in the proper affect, and that those subjects not answer only as their ideal selves) either do not affect many experiments, or they "…affect nearly all of philosophy." However, he misconstrues both versions of the experience machine, his insistence that subjects be rational assumes his conclusion and, if his strictures were taken seriously, they would be disastrous for psychological testing. Lastly, he hints at a use of testing rational subjects that he does not articulate. Weijers, though, is correct about the second horn of his dilemma. Regardless, since many philosophical thought experiments do not require affect to be understood or even concern testing, my two limits do not apply. However, many philosophical thought experiments are subject to these two limitations. In these cases, when we "[…] do not appear" to have any problem understanding them, this is a problem. In fact, we often believe our thoughts are transparent to us when they are not. Therefore, the two limits cited above still apply, regardless of any desire we have to ignore them.[6]

Works Cited

Alexander, Joshua and Jonathan Weinberg. 2007. "Analytic Epistemology and Experimental Philosophy." *Philosophical Compass* 2: 56-80.

Cullen, Shawn. 2010. "Survey Driven Romanticism." *Review of Philosophy and Psychology* 1: 275-296.

De Brigard, Filipe. 2010. "If You Like It, Does It Matter If It Is Real?" *Philosophical Psychology* 23: 43-57.
Griffin, James. 1986. *Well-being: Its Meaning, Measurement and Moral Importance*. New York: Oxford University Press.
Kauppinen, Antti. 2007. "The Rise and Fall of Experimental Philosophy." *Philosophical Explorations* 10: 95-118
Kawall, Jason. 1999. "The Experience Machine and Mental State Theories of Well-Being." *The Journal of Value Inquiry* 33: 381-387.
Milgram, Stanley. 1977. "Subject Reaction: The Neglected Factor in the Ethics of Experimentation." *Hastings Center Report* 7: 19-23.
Nadelhoffer, Thomas and Eddy Nahmias. 2007. "The Past and Future of Experimental Philosophy." *Philosophical Explorations* 10: 123-149.
Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic.
Prinz, Jesse. 2007. *The Emotional Construction of Morals*. Oxford: Oxford University Press.
Silverstein, Matthew. 2000. "In Defense of Happiness: a Response to the Experience Machine." *Social Theory and Practice* 26: 279-300.
Smith, Basil. 2011. "Can We Test the Experience Machine?" *Ethical Perspectives* 18: 29-51.
Sorensen, Roy. 1992. *Thought Experiments*. Oxford: Oxford University Press.
Weijers, Dan. 2012. "We Can Test the Experience Machine; a Reply to Smith." *Ethical Perspectives* 19: 261-268.

NOTES

1. It is worth mentioning that when I ran my own inverted experience machine tests – *nine* of them, in three countries – and added controls (e.g. telling subjects 'imagine that you are a confronted agent', and 'this may be difficult to do'), I got the *opposite* results (Smith 2011, 41). Given my controls, my subjects were *especially rational*. Weijers, then, contradicts his own thesis (i.e. that testing rational subjects is 'an advantage') in disregarding my studies. But my point with these tests was not that they should be accepted, but that, regardless of my controls, these results are only slightly better.

2. Weijers says we should test *rational* subjects, but what can this mean? In practice, for him, such rationality is students answering surveys. Even if students understand what is asked of them, they answer in any position they find themselves, for whatever reasons they find salient (Kauppinen 2007, 104). Such subjects, then, are quite variable. Since his main thesis is that rational subjects yield better results (in some way), Weijers should restrict what rationality amounts to, on pain of offering a vacuous thesis.

3. Antti Kauppinen, in "The Rise and Fall of Experimental Philosophy," points out how, when experimental philosophers give surveys to subjects, these are *not* "straightforward predictions," insofar as they only reveal our "surface intuitions." However, philosophers want our "*robust* intuitions," or those that are "[…] stable under increases in consideration of relevantly similar situations, ideality of circumstances, and understanding of the working of language (centrally, the semantic/pragmatic distinction" (Kauppinen 2007, 110).

4. In "We Can Test the Experience Machine," Weijers interprets my limitation that subjects answer with their "ideal selves" as that they answer "as think they *should*". However, since "ideal selves" or "as we should" mean different things, this is misleading. Weijers seems to vacillate between my notion of "ideal selves," and "should" in a moral sense. However, for sake of brevity let us ignore this and treat "ideal selves" and "should" similarly.

5. Indeed, the *justification* for employing such strictures on rationality applies *equally* not just to the experience machine, but also to cases of helping behaviour, individual or collective responsibility, authority, and a host of philosophical thought experiments, especially those about morality. Therefore, there is no reason to employ these strictures in just one case, yet not apply them to the others.

6. Experimental philosophers like Weijers often seem to want their conclusions for free. Nadelhoffer and Nahmias express this attitude as follows.

> It is incorrect to suggest that experimental philosophers have the burden of demonstrating that the methods of experimental psychology can give us access to philosophically interesting folk intuitions. Rather, the critic has the burden of showing that these intuitions lay forever beyond the reach of experimental philosophy (132).

Is it really illegitimate to ask *how* surveys can yield robust intuitions? As Kauppinen, Cullen and others have argued, experimental philosophers conduct surveys and ask us to believe their results. Given this, they always bear the burden of proof of showing how their surveys yield what they say they do. Indeed, to say anything less is amounts to the *assumption* that their surveys yield the robust intuitions they are after, which is asking us to believe them without evidence (Cullen 2009 282).