

We Can Test the Experience Machine

Response to Basil SMITH “Can We Test the Experience Machine?” *Ethical Perspectives* 18 (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:

“... *certain* 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.” (Smith 2011, p 30, his italics)

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 (Smith 2011, p. 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 that they would actually react if they really found themselves in the hypothesised scenario (Smith 2011, p. 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.*” (Smith 2011, p. 39, his italics).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 (Smith 2011, p. 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 (Smith 2011, p. 46, his italics). 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” (Smith 2011, p. 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:

“...the 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 appropriated confronted agent]. Therefore, in practice, these

thought experiments are impossible to test.” (Smith 2011, p. 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 (Nozick 1974, pp. 42–43). Nozick then asks his readers: “Would you plug in” to an experience machine for the rest of your life? (Nozick 1974, pp. 43). This is clearly a *would* question; one that, on 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 (2010, p. 47) asks “What would you choose?” at the end of all of his scenarios. 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. The stated purpose of De Brigard’s tests was to investigate how informing participants that “you have been living inside an experience machine your entire life... would affect, per se, their judgements on their own happiness or well-being” through studying their responses to his scenarios (De Brigard 2010, p. 46). The stated purpose of Nozick’s experience machine scenario is to demonstrate that “something matters to us in addition to experience” (Nozick 1974, p.44). For both of these stated purposes, 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’ choices to be less rational than they would have otherwise been. However, if we could really know what people think they should choose, then we would have a much better

understanding of our wellbeing-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's 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's 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).ⁱⁱ Therefore, while Smith (2011, p. 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, & Jorgensen 1993, p. 476).ⁱⁱⁱ 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's 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 the kind of thing Smith was getting when he claimed that "...the inverted

experience machine... have a unique set of characteristics that make it impossible to gather the right subjects to test. Therefore, in practice, [it is] impossible to test.” (Smith 2011, p. 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’s 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 a studies 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 don't 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.^{iv} 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:

“... *certain* 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.” (Smith 2011, p 30, his italics)

Of course, there are other reasons why the results of experimental studies may be less reliable than the opinion of philosophers about a particular thought experiment.^v 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 (2011, p. 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's 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.

Dan Weijers

Philosophy Department, Victoria University of Wellington

Works Cited

- De Brigard, F. (2010). If You Like it, Does it Matter if it's Real?, *Philosophical Psychology*, 23(1): 43–57.
- Nozick, R. (1974). *Anarchy, State, and Utopia*, Oxford: Blackwell, 1991.
- Petrinovich, L, P. O'Neill, & M. Jorgensen (1993). An Empirical Study of Moral Intuitions: Toward an Evolutionary Ethics, *Journal of Personality and Social Psychology*, 64(3): 467–478.
- Smith, B. (2011). Can We Test the Experience Machine?, *Ethical Perspectives*, 18(1): 29–51.

Notes

ⁱ 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.

ⁱⁱ 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.

ⁱⁱⁱ "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, & Jorgensen 1993, p. 476).

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

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