Essay

Is the Experience Machine Fatal to a Utilitarian Conception of Justice?

Azel Geist

University of Essex

Abstract

This essay argues that Robert Nozick's (1974) experience machine does not constitute a fatal objection to a utilitarian conception of social justice, even for those utilitarian accounts which employ an experientialist axiology. This is because, in its original form, the experience machine has been experimentally demonstrated to induce materially biased judgements; meanwhile, attempts at modifying it have either failed to induce unambiguous value judgements required to refute experientialism or have generated scenarios which deviate from the original to such an extent that they become tests for subjects' abilities for rational decision-making under uncertainty instead of their value judgements. Furthermore, I find that attempts to save the experience machine by restricting the scope of value judgements considered reliable (the expertise defence) or by re-conceptualising how thought experiments justify ethical claims (the mischaracterization objection) both fail. However, I ultimately concede that the experience machine does provide a weak *pro tanto* reason to reject utilitarian theories of justice, as I cannot rule out the possibility of convincing explanations for the conflicting intuitions generated by de-biased versions of the machine that retain its argumentative force.

Keywords: experience machine, utilitarianism, well-being, social justice

Date of Submission: 24.04.2024 **Date of Acceptance:** 29.08.2024

Introduction

In this essay, I argue that the experience machine is not fatal to a utilitarian conception of justice. Experience machines explicitly target experientialist theories of well-being (hereafter referred to

DOI: https://doi.org/10.5526/esj.365

as "experientialism").[1] As such, an intuitive strategy for defending against such objections is to break the utilitarian dependence on experientialism by adopting an alternative, non-experientialist axiology.

I do not adopt this approach here due to two difficulties. Firstly, there exists a vast array of non-experiential accounts regarding well-being—it is plainly impossible to provide a complete evaluation of each account within the constraints of this essay. The second, more fundamental difficulty with this strategy is that answering the question of whether there is an independently defensible non-experientialist utilitarian conception of justice does not directly bear on the question of whether the experience machine argument succeeds. For whether such a conception may be found would depend on grounds that are considerably independent of the experience machine. As such, I instead restrict the scope of utilitarian theories under consideration to only those that employ an experientialist axiology.

In the following sections, I will demonstrate that it is possible to show that the experience machine is not fatal to utilitarian theories even under such a restricted scope. I first explain how the experience machine has been thought to challenge experientialism, although recent research has discovered that its early formulations tend to be plagued by biases. I then present the experientially identical experience machine, which is often thought to constitute a variant of the original experience machine with its biases neutralised. However, I show that this move suffers from a fatal flaw known as the "freebie problem." Finally, I address two objections: the expertise problem and the mischaracterization problem.

The former is an attempt to sidestep the problem of bias by claiming that only the judgements of persons with a certain "moral expertise" should be accepted when considering the experience machine; the latter seeks to deny the validity of experimental philosophy (ex-phi) approaches in substantiating thought experiments altogether, through re-conceptualising thought experiments in such a way that their argumentative force no longer relies upon the intuitive judgements elicited. I find that neither is sufficient to reinstate the experience machine's status as a fatal challenge to experientialism—consequently, it cannot be used to conclusively reject utilitarian conceptions of justice.

The experience machine & its biases

Utilitarian conceptions of justice rely on a sound theory of well-being to present a persuasive case, for when the central proposition of such theories is that justice simply consists of maximizing the good, failure to completely describe what constitutes the good would undermine them. This is because a utilitarian theory with an incorrect axiology would likely issue erroneous prescriptions

[1] While experience machines have traditionally been thought to only concern hedonistic theories of well-being, their depictions impinge not only on pleasure and pain, as a variety of experiences are possible within the experience machine. Experientialist theories of well-being are those theories which include the claim that well-being may be assessed solely by reference to internal mental experiences.

 Essex Student Journal, 2024, Vol. 16(1)
DOI: https://doi.org/10.5526/esj.365

to sacrifice disproportionate amounts of other intrinsic goods for the sake of the goods recognized in its faulty theory of well-being.

A thought experiment designed to attack this potential vulnerability of utilitarian theories is Nozick's (1974, p. 42) experience machine, which does so by elucidating moral intuitions about what we really value. He tells us to imagine a machine that could stimulate our brains such that we may have "any experience we desired," while we are "floating in a tank, with electrodes attached to our brain." He then gives us a choice—we may either plug into the machine for life (no trial runs) or we may remain unplugged.

If it is the case that we would not wish to plug into the machine—as Nozick believes—then it is argued that this implies that we intuitively value non-experiential goods intrinsically. And since the best explanation for this intuitive valuation is that there are such non-experiential intrinsic goods, we should accept that this is in fact the case. This in turn has the implication that experientialism is false.

This initial version of the experience machine has been severely criticized for inducing judgements that are based on philosophically irrelevant reasons. Löhr (2019, p. 5) lists four such biases that have been identified by prior applications of ex-phi approaches:

1. Self-other bias: persons tend to differ in their judgements to plug in or not depending on whether the subject presented in the scenario is themselves or a stranger.
2. Status quo bias: if persons are portrayed as already inside the experience machine, they are more likely to choose to stay in compared to having the choice to plug in as presented originally.
3. Over-active imagination: descriptions of the experience machine may evoke associations with tropes present in "science fiction horror stories" and lead to the choice being affected by feelings of disgust and horror.
4. Imaginative resistance: stipulations that the machine will function perfectly well and the exclusion of responsibilities to others from consideration may not be internalized successfully.

As such, biases threaten to supplant the proposition that experientialism is wrong as the best explanation of observed judgements not to plug in, attempts have been made to redesign the experience machine to mitigate them. For example, Weijers (2014, pp. 525-526) deals with status quo bias by using a scenario where before the choice to plug in or not is to be made, the subject is stipulated to spend half of their time within the experience machine and the other half outside of it. But when these rectified experience machines are subjected to the same experimental tests, subjects demonstrate conflicting intuitions regarding whether people would choose to plug in (Weijers, 2014). It seems that removing bias fails to fix the experience machine, for the premise that we would choose not to plug in becomes now very dubious.

DOI: https://doi.org/10.5526/esj.365

Moore's heap of filth redux or the experientially identical experience machine

There is another severe problem in the design of the experience machine that has been present from the very beginning: a person can believe that experientialism is false yet still consistently choose to remain in the experience machine. They may do so because of the belief that the valuable experiences obtained in the experience machine are so great in amount that they outweigh the combination of experiential and non-experiential goods that would be gained if they were to remain outside of it. Or perhaps, even if they do not believe the types of goods differ in amount, their intuitions may track a theory of well-being where hedonic goods have a greater weight than non-experiential goods (without dispensing with the intrinsicality of non-experiential goods).[2]

Lin (2016, p. 321) suggests that to eliminate this possibility, the experience machine ought to be altered such that the net value of the experiences obtained is held constant. He terms this the "experientially identical experience machine" (EIEM), which consists of a presentation of two lives, A and B, where the only difference between them is that the subject of A spends their life in the "real world" and the subject of B is plugged into an experience machine. Their internal experiences are identical. If we tend to judge life A to have superior value to life B, this seems to provide better evidence for rejecting experientialism.

However, Weijers (2018, pp. 14 - 19) has found a "freebie problem" in this sort of scheme: rather than successfully plumbing intuitions about the good, the mechanism of EIEMs alters the choice between the competing scenarios into a rational decision under uncertainty. This error arises because of the incorrect assumption that people's decision-making processes may be adequately understood as binary models which accept the inputted intuitive judgements regarding the intrinsicality of a particular good with complete certainty.

How does this occur? If we return to Lin's (2016) scenario, the position of someone judging the value inhering to both lives can be represented by a simple decision model:

> Let p be the probability [0, 1] that only internal experiences have intrinsic value.
>
> Let the utility obtained from internal experiences and living in the real world be non-negative values x & y respectively.
>
> The expected utility for choosing life B is:
>
> $u(B) = x + 0 = x$

[2] This would not be an issue at all if it were true that most people would choose not to plug in, since this would be seen as even more powerful evidence of a plurality of intrinsic goods, as it would be a choice despite the possibility of a greater amount or weight given to experiential goods. This clarifies why the conflicting intuitions observed in de-biased tests of the original experience machine were important motivators for this discovery.

DOI: https://doi.org/10.5526/esj.365

The expected utility for life A is:

$u(A) = x + y * (1 - p) = x + (y - y * p)$

Since p is in the interval [0, 1], $x + (y - y * p)$ is in the interval $[x, x + y]$

Thus, $x + (y - y * p) >= x$

Therefore, $u(A) >= u(B)$

As it has been shown, it is in fact a logical necessity that choosing life A is always the utility maximizing choice regardless of the probability that experientialism is correct—we may say that living in the real world is a "freebie" that ought to be taken just in case that it is valuable.[3]

Misinterpreting the experience machine?

At this point, it appears that despite the flaws in their design being pointed to as the motivation for EIEMs, the original variants fare slightly better—for when their biases can be mitigated as far as reasonably possible, how the conflicting intuitions produced should be explained remains an open question. On the other hand, the freebie problem is baked into the structure of EIEMs, and it is difficult to position a rational decision under uncertainty as anything other than the dominant explanation—the absurd measure of asking participants to become completely certain about their judgements of intrinsic value is obviously impractical.

So perhaps Lin (2016) is right in that experience machines only present at least a weak *pro tanto* reason to reject experientialism, even if he is wrong to ground it on the promise of EIEMs. But the approach that I have hitherto followed to diminish the machine's argumentative force may yet be overturned on the basis that it relies on controversial assumptions underpinning ex-phi methodologies.

The Expertise Defence

The first issue raised by defenders of the experience machine is that thought experiments ought to be understood as attempts to elucidate the intuitions of philosophical experts rather than that of laypersons, because only expert intuitions could be reliably used to infer propositions about what is and is not intrinsically valuable. If so, conflicting intuitions found through the questioning of laypersons may be of little relevance.

Two types of expertise claims undergirding the idea that experts' intuitions have superior reliability are distinguished by Horvath and Koch (2021, pp. 3-4): the mastery model, which asserts philosophers have a superior intuitive grasp of what is intrinsically valuable by virtue of

[3] To point out the obvious, Lin's scenario does show what we have known all along—living in the real world is at least not disvaluable, even though supporters of experientialism would assert that it is valueless in itself.

 Essex Student Journal, 2024, Vol. 16(1)
DOI: https://doi.org/10.5526/esj.365

their extensive training, and the resistance model, which asserts that philosophers have a superior ability to resist biasing conditions.

So, does reality support any of these claims? It appears not. Horvath and Koch (2021, p. 4) describe the mastery model in dismissive terms, as the training received by philosophers is plainly not designed to prepare them to be master intuiters, and the judgements of philosophers in thought experiments are much more like those of laypersons compared to the large differences observed in other fields. The resistance model is also unpromising, as experimental evidence from Löhr (2019) reveals that a significant minority of philosophers gave inconsistent answers to experience machine scenarios with different biasing conditions. Moreover, they were found to give inconsistent justifications for their responses at a similar rate to the lay group.

The Mischaracterization Objection

A radically different line of argument from Horvath (2022) asserts that we have been labouring under a fundamental misunderstanding of the role thought experiments play in arguments. He presents the Deutsch-Cappelen view of thought experiments, which postulates that instead of such experiments generating intuitive judgements directly for further inferences, these judgements are actually inferentially justified by giving independent arguments (and thus are in fact justified by intuition).

If we accept this view, we can read Nozick (1974, p. 43) as giving three reasons for the judgement not to plug in. He says that: a. "we want to do certain things", rather just have the experience of doing them; b. "we want to be a certain way", claiming that our sense of identity depends on contact with reality; and c. plugging in would fail to allow one to be in contact with "deeper reality" and thus result in losing the possibility of deeper meaning.

Yet, these reasons are not sufficiently convincing to be fatal to experientialism. For example, a hedonist can appeal to the paradox of hedonism to explain our desire for actual doing as stemming from the fact that direct pursuit of experiences (including pleasure) is seldom very successful in obtaining said experiences (Crisp, 2006, p. 637). It is also not clear that our identities are formed by our qualities that manifest in the "real world" rather than those manifesting from experiences. The assertion that losing contact with the "real world" would entail a loss of meaning is poorly supported—simply assuming that the metaphysics of meaning would preclude it from permeating from the "real world" into the experience machine, perhaps by inhering in experiences themselves, is unsound.

Conclusion

I have found that the extant methods of modifying the experience machine to immunize it against bias lead only to the production of conflicting intuitions regarding well-being, or a transformation into a rational choice problem under uncertainty that is irrelevant to the axiological considerations of interest. Furthermore, given that the expertise defence does not successfully

DOI: https://doi.org/10.5526/esj.365

challenge the empirical approach adopted by ex-phi, while the mischaracterization objection, even if successful on its own terms, does not lead us to a formulation of the experience machine that can reject experientialism unequivocally, I conclude that it is not fatal to utilitarian conceptions of justice. Crucially, I have thus also shown that this conclusion can be reached without the need to decouple utilitarian theory from an experientialist axiology—relying on the internal flaws of the experience machine is sufficient.

However, since it may be possible to explain away the conflicting intuitions generated by de-biased forms of the experience machine, I concede at this time that the experience machine provides a weak *pro tanto* reason to reject utilitarian conceptions of justice. This also suggests that in order to produce a more definitive conclusion, future work should focus on evaluating candidate explanations for said conflicting intuitions.

References

Crisp, R. (2006) 'Hedonism reconsidered', *Philosophy and Phenomenological Research*, 73(3), pp. 619-645. Available at: https://doi.org/10.1111/j.1933-1592.2006.tb00551.x

Horvath, J. (2022) 'Mischaracterization reconsidered', *Inquiry*, pp. 1-40. Available at: https://doi.org/10.1080/0020174X.2021.2019894

Horvath, J., Koch, S. (2021) 'Experimental philosophy and the method of cases', *Philosophy Compass*, 16(1), article number e12716. Available at: https://doi.org/10.1111/phc3.12716

Lin, E. (2016) 'How to use the experience machine', *Utilitas*, 28(3), pp. 314–332. Available at: https://doi.org/10.1017/S0953820815000424

Löhr, G. (2019) 'The experience machine and the expertise defense', *Philosophical Psychology*, 32(2), pp. 257-273. Available at: https://doi.org/10.1080/09515089.2018.1540775

Nozick, R. (1974) *Anarchy, state, and utopia*. New York: Basic Books.

Weijers, D.M. (2014) 'Nozick's experience machine is dead, long live the experience machine!', *Philosophical Psychology*, 27(4), pp. 513-535. Available at: https://doi.org/10.1080/09515089.2012.757889

Weijers, D.M. (2018) 'The freebie problem: A pervasive flaw in how we work out what has value', *Philosophy Research Seminar Series*, University of Auckland, Auckland, New Zealand, 17 October 2018. Available at: https://hdl.handle.net/10289/12999

DOI: https://doi.org/10.5526/esj.365

Copyright Statement

This article is CC BY (Azel Geist) Essex Student Journal, 2024, Vol. 16(1)
DOI: https://doi.org/10.5526/esj.365