The Desire Machine

0. Introduction

The experience machine poses the most important problem for hedonist theories of well-being. Nozick imagines neuroscientists who create a machine that provides you with the simulation of any experience that you want.[[1]](#footnote-1) It can provide a large amount of pleasure, even though you do not take pleasure in anything in the external world. Most people have the intuition that plugging in to the experience machine would not be good for you.[[2]](#footnote-2) This is contrary to hedonism.

But neuroscientists have recently created another machine that achieves a similar result by different means: the desire machine. Upon entering this machine, your desires are altered through some minor neurosurgery. In particular, the machine causes you to desire everything that actually happens. When you come out of the machine and continue to interact with the real world, your experiences of that world are very different and uniformly positive, because you have a pro-attitude to everything that actually happens. Before entering the desire machine, you often had unsatisfied wants that reduced your quality of life, but with this stunning new treatment, all sources of frustration are gone. Since you approve of everything that actually happens, you lose any motivation to act or change how things are, and you slide into a pitiable lethargy.

The experience machine constructs a simulated world that matches your preexisting desires. The desire machine reconstructs your conative state to match the preexisting world. The experience machine is *the* problem for hedonist theories of well-being. The desire machine is *the* problem for desire satisfactionist theories of well-being. One experiences pleasure but one’s desires are not actually satisfied in the experience machine, so desire satisfactionism (correctly) entails that one should not enter and hedonism (incorrectly) entails that one should. But upon entering the desire machine, one’s desires are actually satisfied and one thereby experiences pleasure, so hedonism and desire satisfactionism both (incorrectly) entail that one should enter.

In this paper, I consider the question of whether one should enter the desire machine. The sense of “should” I am interested in is *prudential* and *objective*: when we think about whether an agent should enter the desire machine, or whether it would be rational for that agent to enter the desire machine (I will use these locutions interchangeably), all we should consider is the decision’s impact on the agent’s well-being, holding everything else fixed. Likewise, we should consider whether entering the desire machine would actually be good for an agent, not whether the agent’s evidence indicates that entering the machine would be good for her.

With this sense of “should” specified, it seems clear that one should not enter the desire machine: entering the machine would not be good for one.[[3]](#footnote-3) Yet, desire satisfactionism entails that one should enter. According to that theory, having satisfied desires is what makes one well-off. There are two ways to have satisfied desires: make the world conform to your desires or make your desires conform to the world. The desire machine takes the latter route, leaving one with many satisfied desires and, according to desire satisfactionism, well-off. I will consider how desire satisfactionists might avoid this result. First, I discuss the phenomenon of change of desire, of which the desire machine is a special case. Second, I consider coherence-based norms of rational desire change. Finally, I argue that introducing a substantive account of fitting desire is the only plausible solution, but that this response requires abandoning pure subjectivism about well-being.

1. The Problem: Change of Desire

Our desires change a lot over time. This much is empirical fact. As they age, people gain and lose desires that are appropriate to their stage of life. Children lose desires to be astronauts when they grow up. Young parents gain many desires that they did not have as childless young adults. Older adults lose desires they had when they were younger. Gaining or losing a capability changes what desires people have. There is much discussion of the empirical phenomenon of adaptive preferences, which occurs when someone changes their desires when their old desires are impossible to satisfy.[[4]](#footnote-4)

Given that our desires change over time, the question arises of whether this is a rationally evaluable phenomenon. Could one have reasons to change what one desires, or is change of desire something that just happens to people? Compare desire to belief. Our beliefs, too, clearly change over time. And change of belief clearly is rationally evaluable—epistemologists debate how to rationally change your beliefs in response to new evidence. I will assume that change of desire is also rationally evaluable.

The question then becomes: what are the norms for change of desire? There are two incompatible lines of reasoning that both seem attractive.[[5]](#footnote-5) The first has some appeal but ultimately does not work out. The second is what leads us straight into the desire machine.

The first option is that changes of desire are evaluated with respect to the desires the agent has before the change. This thought has a clear motivation. Suppose that you are making a decision that would change what desires you have. This, like any other decision, should be evaluated using expected utility theory. Utilities must be assigned to states using your current desires and probabilities must be assigned to those states using your current beliefs. In particular, states in which you do not have a desire D are assigned a utility by D and your other desires. But what expected utility should I assign to the state in which I have coffee in the morning but do not want coffee in the morning, given that I want coffee in the morning now?

It seems that low expected utility must be assigned to states in which you have different desires. A desire that P is the kind of thing that has a tendency to cause you to bring about P. But if you do not desire P, then P is less likely to come about since you would no longer be trying to bring it about. From the perspective of your current desires, then, the state in which you lose a desire should be negatively evaluated because some things that you desire now would be less likely to come about in the state where you do not desire them. If the above line of reasoning is adopted, it looks like just about any change of desire would come out as irrational. The problem is that even the seemingly natural and appropriate changes of desire that take place over the course of a lifetime are in fact a huge mistake: we should have never abandoned the desires that we had when we were children.[[6]](#footnote-6)

The second option is that changes of desire are evaluated with respect to the agent’s desires after the change. This does not have overly conservative implications like the first option, but at present the thought is underspecified. Suppose that the agent can acquire one of two desires, D1 and D2. What should the agent do? All things equal, the state in which an agent has D1 is assigned higher utility by D1 than the state in which the agent has D2, and the state in which the agent has D2 is assigned higher utility by D2 than the state in which the agent has D1. So the choice between D1 and D2 is underdetermined. But there is something that could decide between D1 and D2. If one is a more realistic or feasible desire (i.e., it is easier to bring about), then arguably it should be formed. If D1 is a desire to go to Mars, and D2 is the desire to take up woodworking, then it would seem that the agent has reason to form D2 and not D1 because D1 would be very hard to satisfy. If the agent forms D2, it is more likely that the agent will be satisfied, and this seems to break the tie in favor of D2.[[7]](#footnote-7)

This starts us down the slippery slope to the desire machine. If desires that are easy to satisfy are better, then desires that are already satisfied are the best. If you could come to want everything that actually happens, this would be perfectly rational on the view that changes of desire are evaluated by the desires one has after the change.

The desire machine is only the extreme case of a common phenomenon: change of desire.[[8]](#footnote-8) The phenomenon of change of desire poses problems for desire satisfactionists. They think that a person’s good is grounded in the satisfaction of their desires, but when we move to ask whether changes of desires (the ground of a person’s good) are good or bad for a person, there appears to be nothing besides the amount of desire satisfactions a person has that could account for the rationality of desire change. This is what leads us to the desire machine, which gives one the greatest level of desire satisfaction possible. Desire satisfactionists need to develop an account of this phenomenon that allows them to vindicate the changes of desire we take to be appropriate, but one that also allows us to reject the implication that one should enter the desire machine. I will discuss two possible accounts.

2. Conative Coherence

Humeanism is the view of desire on which there is no standard of correctness desires must answer to, external to the agent’s perspective. In this respect, desire is unlike belief. Hume himself infamously eschewed even coherence-based reasons to have one desire rather than another: “Tis as little contrary to reason to prefer even my own acknowledg’d lesser good to my greater, and have a more ardent affection for the former than for the latter.”[[9]](#footnote-9)

More moderate Humeans have posited that the only norms of rational desire change derive from the coherence among one’s desires.[[10]](#footnote-10) Here are some examples of plausible coherence-based constraints on desire:[[11]](#footnote-11)

1. One should not desire P and desire ~P.
2. One should not desire something that one knows is impossible.[[12]](#footnote-12)
3. If one desires P and one desires Q, then one should desire P and Q.
4. A person should desire to take the means to her desired ends.
5. One’s object-level and meta-desires should be in alignment: one should not desire P and desire that one does not desire P.

Desire satisfactionists could argue that change of desire is rational iff it conforms to these and other similar constraints.[[13]](#footnote-13) In particular, they could argue that a change in desire is rational iff it moves one closer to conative reflective equilibrium. Conative reflective equilibrium is a state which mirrors its cognitive counterpart, described by Goodman, Quine, Rawls, Lewis and many others. It is a state in which one’s desires are maximally coherent and free from internal tensions.

Could desire satisfactionists use the notion of conative coherence to avoid the implication that one should enter the desire machine? Though there are no *substantive* constraints concerning which desires one should have, form or lose (this idea is at the very core of subjectivism about well-being), there are constraints deriving from the coherence among one’s desires. It is rational to change one’s desires in ways that move one’s overall web of desire towards reflective equilibrium. These changes are essentially modest, and the rationality of each of the changes has its basis in one’s current desires. Entering the desire machine is a much more substantial change than the incremental changes that the coherence approach recommends. The desires one would get from the desire machine, though they may be in conative reflective equilibrium, would not be a coherent extension of one’s current desires, and as such one should not enter the desire machine.

I have three concerns about this proposal. First, the analogy between cognitive and conative reflective equilibrium is not perfect. On most models of cognitive reflective equilibrium, the process is one of revising one’s beliefs in light of new information about the external world that one gains through a learning experience. Observations can be rejected if they come into acute contrast with one’s previous beliefs, but there is at least some source of new data driving how one’s web of belief evolves over time. It is unclear what the analogous feature of a conative model of reflective equilibrium would be. The input one gets from observation is what ties models of reflective equilibrium to the world. Without this constraint, it is not clear how a conative model of reflective equilibrium would work at all, since it would have a quite substantial free parameter.

Dorsey (2010, 547-549) develops a view of conative coherence that includes an account of what he calls “value data” that constrain conative reflective equilibrium. His account differs somewhat from the one I just sketched. He construes the things which are to be brought into conative coherence as evaluative beliefs, not desires, and his aim is to argue that the evaluative beliefs that survive a process of reflective equilibrium are those that are relevant to the individual’s well-being. The value data are judgments about the quality of a life based on actually experiencing living that life. Unfortunately, Dorsey’s value data are not the kind of thing that could prevent one from choosing the desire machine—a life in the desire machine would certainly be judged to be better by the person who lived it than an ordinary life, but this does not mean that it is better. More importantly, each person has access to exactly one value datum, which is not sufficient to guide rational changes in one’s desires over time.

Second, coherence alone cannot justify the changes in desire we take to be appropriate. My current desires are hardly a coherent extension of the desires I had when I was five years old—the change is just too substantial for the former to be a more coherent version of the latter. This large change in desire is not a failure of rationality on my part; indeed, it is perfectly appropriate. The coherence approach is unintuitive in other respects, since, in general, coherence-based standards are too easy to satisfy. Someone who has no desires has a perfectly coherent set of desires, but we rightly judge that it would be better for a person with no desires to gain some desires. Likewise, defective desires that fit into a defective web of desires can be justified by those other desires. Stalin’s formation of a desire to send his enemies to Siberia is appropriate, given the rest of his desires.[[14]](#footnote-14) Coherence alone, even if it solves the problem related to the desire machine, is inconsistent with our intuitions and creates problems elsewhere.

Third, there are two distinct questions we can ask about both the desire and the experience machines: ought one to enter it, and is one’s well-being higher in the state in which one enters it? The coherence-based approach only allows us to say that one ought not enter the desire machine, not that one’s level of well-being is lower in the desire machine than it is on the outside. One ought not enter it because doing so would leave one with desires that are not a coherent extension of one’s current desires, but a life that is lived entirely within the desire machine has a higher level of well-being than one lived entirely on the outside because more of one’s desires are satisfied in the former than in the latter.[[15]](#footnote-15) Driving a wedge between rational requirements and well-being, as the coherence-based approach does, is itself highly problematic. What, especially from a subjectivist perspective, could make it the case that one prudentially ought to do something except for considerations related to one’s own well-being? I submit that we should aim for a more robust response to the thought experiment, one that also allows us to say that one’s level of well-being is not higher in the desire machine, not just that one should not enter the machine.

3. Fitting Desire

I want to suggest a modification of desire satisfactionism that allows us to provide a more robust response to the desire machine. The key change in perspective that this approach requires is thinking of desires just like other intentional mental states, like belief, fear, admiration, envy, etc. These other states have *standards of correctness*. To use an increasingly popular concept, they can be fitting or unfitting to their objects. Fear of X is fitting if X could harm you, admiration of Y is fitting if Y is admirable, and belief in Z is fitting if Z is true. These attitudes have standards of fit because they are representational: they can represent their objects either correctly or incorrectly. It would be surprising if desire lacked a standard of correctness, though all of the other attitudes which are similar to it have standards of correctness. Desires, just like beliefs, have contents and represent their content as being a certain way: desires represent their contents as being good. This is a marked departure from a Humean view of desire on which any desire is just as correct as any other.

Here is the final piece of the puzzle: a satisfied desire gives you more well-being the more fitting it is. On this view, it is not necessarily the case that the satisfaction of unfitting desires gives you no well-being. This is because fittingness is a gradable property and it may be that no desires are maximally fitting or unfitting, though some are more fitting than others. The more fitting a satisfied desire is (i.e., the more desirable its object is), the more well-being one gets from the satisfied desire.

One should not enter the desire machine because many of the desires that one would get from the desire machine would be unfitting. Since one’s well-being is a fittingness-weighted sum of one’s satisfied desires, even though one has more satisfied desires as a result of the desire machine, one’s well-being is lower because the fittingness-weighted sum of one’s satisfied desires is lower. It is clear that desiring everything that actually happens does not increase one’s level of well-being after one’s desires are adjusted for fittingness. One will come to desire all of the injustice that occurs in the world, for example, upon entering the desire machine, but this is not the kind of desire which could give one well-being even if satisfied. More generally, one will become extremely apathetic upon entering the desire machine—never forming a desire for things to be different from how they actually are. As such, one will never undertake any projects or strive towards any goals. These desires for stasis and complacency are not fitting. That is why one should not enter the desire machine.

It is important to note that this response constitutes a departure from pure subjectivism. By introducing the notion of desirability—a property that things have when it is fitting for one to desire them—we are committed to a standard by which desires can be assessed that is independent of an agent’s actual desires. However, this is only a partial departure, for the standard in question need not be independent of all facts about the agent. Introducing the notion of fittingness is consistent with the agent-relativity of the evaluative properties in question: something can be desirable from my perspective (fitting for me to desire), but not desirable from your perspective (not fitting for you to desire). For example, it is fitting for my dog to desire to eat dogfood but it is not fitting for me to desire to eat dogfood, so the dogfood is desirable from my dog’s perspective but not desirable from my perspective—desirability is agent-relative.[[16]](#footnote-16) The theory that results is “objective” in the sense that there is something independent of an agent’s actual desires by which they can be assessed, and which partially determines how much well-being the agent enjoys, but not “objective” in the sense that the standard in question is the same for all agents because it is a mind-independent property of the objects of desire. This, I contend, is a level of objectivity that desire satisfactionists should be willing to adopt in order to avoid the catastrophic implication that one should enter the desire machine.

I have argued that desire satisfactionism faces a problem analogous to the main problem facing hedonism. Indeed, entering the desire machine is arguably *worse* than entering the experience machine: the person in the experience machine is pitiable because of their circumstances, in which every meaningful thing in their life is based on false pretenses, but the pitiable thing about the person in the desire machine is the kind of person they have become as a result of entering the machine. Our intuitions are sharply opposed to the claim that entering the desire machine is good for one, and I argued that desire satisfactionists do not have the resources to resist this conclusion. They may try to offer an error theory about our intuitions and embrace the result that entering the machine is good for one after all, like some hedonists have in the case of the experience machine, but ultimately, abandoning pure subjectivism may be a more fitting response.[[17]](#footnote-17)

Bibliography:

Barnes, Elizabeth. 2016. “Taking Their Word for It.” In *The Minority Body: A Theory of Disability*, edited by Elizabeth Barnes, 0. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198732587.003.0005>.

Bovens, Luc. 1992. “Sour Grapes and Character Planning.” *Journal of Philosophy* 89 (2): 57–78. <https://doi.org/10.2307/2027152>.

Bykvist, K. 2006. “Prudence for Changing Selves.” *Utilitas*, 18(3), 264-283. doi:10.1017/S0953820806002032

De Brigard, Felipe. 2010. “If You Like It, Does It Matter if It’s Real?” *Philosophical Psychology* 23 (1):43-57.

Dorsey, Dale. 2006. “A Coherence Theory of Truth in Ethics.” *Philosophical Studies* 127 (3): 493–523.

———. 2010. “Preferences, Welfare, and the Status-Quo Bias.” *Australasian Journal of Philosophy* 88 (3): 535–54. <https://doi.org/10.1080/00048400903225130>.

———. 2013. “Desire-Satisfaction and Welfare as Temporal.” *Ethical Theory and Moral*

*Practice* 16 (1): 151–71. <https://doi.org/10.1007/s10677-011-9315-6>.

———. 2017. “Adaptive Preferences Are a Red Herring.” *Journal of the American Philosophical Association* 3 (4): 465–84. <https://doi.org/10.1017/apa.2017.36>.

Elster, Jon. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139171694>.

Heathwood, Chris. 2006. Desire satisfactionism and hedonism. *Philosophical Studies* 128 (3): 539-563.

Hume, David. 2003. *A Treatise of Human Nature*. New edition. Mineola, N.Y: Dover

Publications.

Lin, Eden. 2016. “How to Use the Experience Machine.” *Utilitas* 28 (3): 314–32. <https://doi.org/10.1017/S0953820815000424>.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. Basic Books.

———. 1991. *The Nature of Rationality*. Princeton University Press.

Pettigrew, Richard. 2019. *Choosing for Changing Selves*. Oxford University Press.

1. Nozick (1974, 42-45). [↑](#footnote-ref-1)
2. For empirical work on intuitions regarding the experience machine, see De Brigard (2010). [↑](#footnote-ref-2)
3. These two questions are conceptually distinct. It is conceivable that one prudentially shouldn’t do something even though it would make one better-off, or that there are things one prudentially should do even if they don’t make one better-off. This is not my view; I assume throughout that actions that make one better-off are always actions one prudentially should perform and vice versa. If our best account of the desire machine required giving different answers to these questions, separating well-being from the prudential ought, that would itself be an interesting philosophical discovery (more on this in section 2). [↑](#footnote-ref-3)
4. For discussions of this phenomenon, see Barnes (2016), Dorsey (2017), Elster (1983), Bovens (1992). [↑](#footnote-ref-4)
5. Pettigrew (2019) addresses a similar problem. [↑](#footnote-ref-5)
6. In addition, some views about the temporal structure of well-being (e.g. Heathwood 2006) suggest that one does not benefit from P if P obtains at only times when one does not desire P. If this is the case, even if P happened to come about after you stopped desiring it you would not benefit, a reason not to lose the desire. [↑](#footnote-ref-6)
7. Bykvist (2006, 279-280) endorses a similar principle (though not one that concerns feasibility): if your possible future self A wants the life that it will lead more than your possible future self B wants the life that it will lead, then life A is better for you than life B. [↑](#footnote-ref-7)
8. The experience machine is likewise only the extreme case of a common phenomenon: pleasure based on a false belief. [↑](#footnote-ref-8)
9. Hume (2003, 2.3.3.6). [↑](#footnote-ref-9)
10. Nozick (1991, 139-151). [↑](#footnote-ref-10)
11. There are also coherence-based constraints on preferences, the comparative form of desire, such as transitivity, asymmetry, etc. [↑](#footnote-ref-11)
12. Nozick (1991, 144). [↑](#footnote-ref-12)
13. Dorsey (2006), (2010) and Bovens (1992) develop such approaches. [↑](#footnote-ref-13)
14. This is a conative analogue of the “coherent fairy tale” problem for coherentist theories in epistemology. [↑](#footnote-ref-14)
15. Some recent work on the experience machine (e.g., Lin 2016) has urged that formulations of the thought experiment should focus on comparing lives inside and outside the experience machine in terms of their levels of well-being, rather than asking whether one would or should enter the machine. [↑](#footnote-ref-15)
16. Here again, desire is not an outlier. Many (perhaps all) other attitudes with a standard of fit correspond to agent-relative evaluative properties. Roughly, it is fitting for a wrongdoer to regret his wrongdoing but it is not fitting for others to do so. It is fitting to blame someone only if one has standing to blame, but not fitting for those without standing to do so. It is fitting for a mouse to fear my cat but not fitting for me to fear my cat. [Redacted]. Thanks to two referees at *Analysis* for encouraging me to clarify this point. [↑](#footnote-ref-16)
17. [Acknowledgements Redacted] [↑](#footnote-ref-17)