Article:
Hayward, M.K. (2020) Utility cascades. Analysis, 80 (3). pp. 433-442. ISSN 0003-2638

https://doi.org/10.1093/analys/anaa011

Utility Cascades

Forthcoming in *Analysis*

MAX KHAN HAYWARD
University of Sheffield
mh3173@columbia.edu

Abstract

Utility Cascades occur when a utilitarian's reduction of support for an intervention reduces the effectiveness of that intervention, leading the utilitarian to further reduce support, thereby further undermining effectiveness, and so on, in a negative spiral. This paper illustrates the mechanisms by which utility cascades occur, and then draws out the theoretical and practical implications. Theoretically, utility cascades provide an argument that the utilitarian agent should sometimes either ignore evidence about effectiveness, or fail to apportion support to effectiveness. Practically, utility cascades call upon utilitarians to rethink their relationship with the social movement known as Effective Altruism, which insists on the importance of seeking and being guided by evidence concerning effectiveness. This has particular implications for the "institutional critique" of Effective Altruism, which holds that Effective Altruists undervalue political and systemic reforms. The problem of utility cascades undermines the Effective Altruist response to the institutional critique.

1. Introduction

On the basis of a series of ingenious (and hilarious) thought experiments, Ryan Doody (MS) argues that a distinctive flaw of certain non-utilitarian moral theories is that they licence *ostriching*. By these theories' own lights, agents will do better if they blind themselves to various morally salient features of their situations – sticking their heads in the sand, as it were (although, as Doody wryly notes, this is not actually an accurate description of ostrich behaviour). He assumes that utilitarian theories do *not* licence ostriching, and that this constitutes a significant point in favour of utilitarianism.

But this assumption is too quick. There are cases where, by the act-utilitarian's own lights, she would do better if she stuck her head in the sand and refused to update her judgements in the light of new evidence. This is especially salient where *utility cascades* arise. These occur when ongoing rational updating of judgements concerning the effectiveness of an intervention causes a utilitarian to push a situation further and further away from the antecedently optimific outcome. Utility cascades are, I suggest, pervasive in our world.

This is a serious theoretical issue for act-utilitarianism (henceforth, my argument will focus specifically on act-utilitarianism). Utilitarians need to clarify how they understand the relative priority of moral and epistemic normativity. It is also a serious *practical* issue. Utilitarians must rethink their alliance with 'effective altruism,' a movement born in large part from the utilitarian tradition. Effective altruists insist that agents must always act in the light of the best evidence concerning the effectiveness of an intervention - that is, they should never ostrich. Once the significance of utility cascades is appreciated, it might turn out that the best practical (and political) implementation of utilitarianism looks quite different from what effective altruists propose.

Whether all of this constitutes an argument against utilitarianism, or simply illustrates the under-appreciated – and potentially radical – implications of the doctrine, I leave to the reader to judge.

2. Hard times in Altruistan

Bill tries to be a very good chap. He is an act-utilitarian, and not just any old akratic or irrational act-utilitarian. Bill is an Effective Altruist, and resolutely directs his waking hours (not just 80,000 of them!) to the maximisation of Utility. Well, that is what he would like to do. But the best he can manage, as a sadly limited being, is performing the actions, out of those available to him, with the highest expected utility. And so he meticulously updates his expectations concerning utility in the light of the best evidence available to him, and calibrates his actions accordingly. He is not just a good chap, but also (Bill would say: '*therefore*') an epistemically *rational* chap.

Bill has considerable monetary resources, and devotes these to supporting various worthy initiatives. He apportions his support in the following way. First, he assesses two things – both how much utility the initiative will secure if it succeeds, and how likely it is to succeed on the supposition that he supports it. This allows him to assign an expected utility – or *effectiveness* – score to the initiative. Then he compares the initiative to other initiatives, and supports highly effective initiatives in proportion to their effectiveness score.

One initiative distributes a vaccine – effectanol[1] – with the aim of preventing the spread of an agonising infectious disease in Western Altruistan. The vaccine does not guarantee protection from the disease. But Bill has been informed that it significantly *reduces* the likelihood of infection. The vaccinated are only 20% likely to catch the disease. If they are not infected, they will not transmit the disease to others. So the more people who receive the vaccination, the more the spread of the disease will slow. Bill calculates that effectanol is likely, given his backing, to reach many people, and thus prevent a great deal of suffering. Good stuff – backing effectanol is both utility-promoting if it succeeds, and quite likely to succeed. It is *effective*. Bill decides to donate $10,000 a month. The rate of new infections starts to slow.

That was back in June, though. Come July, Bill receives disturbing news. The initial trials were misleading. Effectanol is not quite as reliable as once thought. In fact, the vaccinated are 30% likely to fall ill. Bill revises the effectiveness score he had given to the effactanol initiative, and, regretfully, decides only to support it to the tune of $8,000 a year. There are, after all, other initiatives calling for his attention, and his principles tell him to distribute his resources in proportion to effectiveness scores. He sends the balance from the $10,000 to a program which distributes mosquito nets in sub-Saharan Africa – an old standby.

August rolls around. It turns out that the loss of $2,000 dollars makes a big difference in Altruistan. Without Bill's backing for the effectanol initiative, vaccination rates are struggling to keep up with

[1] I follow Doody's protocols for the nomenclature of fictitious substances.

infection rates. Although vaccinated individuals will not spread the disease, there are now so many infected people that the marginal value to the *unvaccinated* of each new vaccination has declined – they're increasingly likely to catch the disease from other people. In other words, the effectiveness score of the effactanol initiative is declining. Even if Bill were to bring his donation back up to $10,000 a month, his intervention would have a lower effectiveness than it did in June, since so many more people now have the disease. He could not in good conscience commit more than $10,000, since there are still many other causes demanding his attention. In fact, even $8,000 seems excessive given current odds – mosquito nets are so reliable. He cuts his donation to $4,000 a month.

September once offered a respite from the blazing heat of Altruistani summers, but this year there is no respite. Vaccination rates have been too low, and infection is now extremely widespread. Bill calculates, regretfully, that the effectanol intervention now has little chance of offering population-level effects – its only benefit will be to those who are vaccinated. And the vaccine is only 70% likely to protect even them. The September effectiveness score for effectanol, it turns out, is far lower than it was back in June – putting it well below other initiatives on Bill's radar. Bill, never one for sentiment or sunk costs, ceases his donations. Bill's donations have done little good, and misery reigns in Altruistan.

Bill has been forced, by his own principles, to abandon the initiative in the light of the changing situation. But much of the change in the situation is attributable to his own actions! And his actions were all performed in the light of his utilitarian principles and the best evidence. Bill has been caught in a *utility cascade*. Had he never received that initial revision of the reliability of effectanol – or had he ignored it – he would not have reduced his donations. If he had not reduced his donations, there would not have been a spike in infections. If there had not been a spike in infections, the effectiveness score of the effectanol intervention would not have dropped. If the effectiveness score of the effectanol intervention had not dropped, Bill would not have reduced his donations even further. And so on.

3. Stormy Weather

As I mentioned before, Bill controls significant monetary resources. But he is no shadowy philanthropist or anonymous donor. Bill is a leading figure in the global Effective Altruist movement – Alt-F as they call themselves – a network containing a bevy of philanthropically inclined technology billionaires who live by (what they take to be) the tenets of Act Utilitarianism.

There is wide consensus in Alt-F concerning the significance of climate change for global utility. It's understood that, as global warming approaches and then exceeds a 2°C temperature rise globally, changes that precipitate great suffering will become more and more likely. Huge tracts of land will become inhospitable to agriculture and some will even become too hot for habitation; rising seas will submerge populous cities. Some locales will benefit from rising temperatures (Michigan will finally be habitable in the winter; the English wine industry will celebrate), but the mass movements of people towards the poles and away from the coasts will likely cause global strife.

The Alt-F movement backs two kinds of initiative to confront the misery that will stem from climate change. The first are *preventative* initiatives. These aim to reduce the rate of global warming – by taxing carbon emissions, for example, or supporting research to make solar panels cheaper. The second are *mitigating* initiatives – these reduce the misery that warming will bring about, if it happens. These include research into drought-hardy crops, the building of sea-barriers, preparation for wildfires, and programs to facilitate mass migration away from devastated regions. Of course, the more efficacious prevention turns out to be, the less mitigation will be needed; the less efficacious prevention is, the more mitigation will be called for.

Bill understands that preventative strategies are fairly likely to work in minimising global temperature rises, given sufficient resources. They thus score quite highly on effectiveness. So Bill devotes the bulk of his resources to prevention, and smaller amounts to mitigation. But then the bad news comes in. Several of Bill's favourite preventative projects turn out to have been oversold. Their chances of significantly preventing warming, while still good, are not quite what Bill thought. This lowers their effectiveness scores in Bill's eyes. Correspondingly, it raises the effectiveness scores of mitigation strategies, since mitigation is more likely to be needed. Bill dutifully transfers some of his backing from prevention to mitigation.

This time, the utility cascade isn't limited to Bill. Other Alt-Fers have heard the news that prominent prevention strategies have been overrated. And they've been following Bill's philanthropy. With Bill's backing reduced, they realise that prevention is *even less* likely to suffice than initially expected. So they also shift resources from prevention to mitigation.[2] Bill notices the increased loss of support for prevention, and dutifully ups his support for mitigation – at the expense of prevention. And so on – you get the picture. Soon, the majority of investment has shifted to mitigation – for warming over 2°C is now highly likely to occur.[3]

In this case, the utility cascade has two additional elements. First, the effectiveness of prevention and mitigation are not independent. The lower the expected utility of prevention, the higher the expected utility of mitigation. And, secondly, it is not only Bill's actions that are governed by changing efficiency scores. The other Alt-Fers similarly direct their support in proportion to efficiency scores. As in the effectanol case, the efficiency of prevention strategies is partly dependent on their degree of support. So the initial downgrading of prevention's effectiveness score leads, via the response of Bill and other Alt-Fers, to a spiralling reduction in the effectiveness of prevention, and, eventually, to the collapse of preventative strategies.

[2] In cases of *decreasing* marginal utility, each donation is worth more the *fewer* other donations there are, and so we would expect withdrawn donations to eventually be replaced by other effective altruists as the expected utility of donations goes up. What I am suggesting is that climate change intervention is - at least up to a point - a case of *increasing* marginal utility. The more people back it, the more expected value each intervention has.

[3] Such reasoning is not limited to hypothetical utilitarians dwelling in the pages of academic philosophy journals. Writing in *The New Yorker,* Jonathan Franzen asserts that 'All-out war on climate change made sense only as long as it was winnable. Once you accept that we've lost it, other kinds of action take on greater meaning. Preparing for fires and floods and refugees is a directly pertinent example.' (Franzen 2019)

4. Diagnosis

Bill apportions his support based on effectiveness scores. But the effectiveness scores of the interventions considered are not independent of his support! So his support is both *determined* by effectiveness calculations, and partially *determinative of* the effectiveness of an intervention.

Of course, Bill recognises the relationship between his support and the effectiveness of an intervention. This is why, when he calculates effectiveness ratings, what he attempts to determine is the expected utility of an intervention *given that he supports it*. But this calculation is not straightforward. First, his support is not binary, but a matter of degree. Second, and vitally, whether or not he supports – and continues to support – an initiative is not really up to Bill. Some people follow the money; Bill follows effectiveness ratings. So long as Bill is consistent in his principles, he *cannot* support any initiative at any degree beyond that determined by comparative effectiveness calculations. His future decisions are not really under the control of his present self – they answer only to the evolving dictates of the effectiveness calculus.

Bill may thus try to take his susceptibility to utility cascades into account, treating his future decisions as simply another part of the situation to be assessed. This might mean assigning a lower effectiveness score to any intervention that is susceptible to cascades. But this does not solve the problem. Whatever effectiveness score Bill gives to the effectanol initiative, he must still decide whether to back it or not. If he backs it, he will be vulnerable to utility cascades. He may *continue* to take the possibility of cascades into account as he updates effectiveness scores, so that once a cascade becomes sufficiently likely, he will immediately pull all his support – as he might have done during July or August in Altruistan. But this would still be a utility cascade – merely a faster one. Bill has tanked an initiative that might have borne fruit, wasting his initial investment.

Recognising that his backing contributes to the *future* effectiveness of the initiative, Bill might consider continuing to donate $10,000 in July, *despite* the bad news about effectanol's efficacy. But that is only rational if he will *keep* donating in August. Yet in August, continued donations will only be rational if he will *keep* donating in September…and so on. And Bill knows his future self will only do what is rational. Bill cannot decide whether to back the initiative until he knows it is rational, but he cannot know it *is* rational until he knows whether his future self – in much the same situation – deems it rational!

This brings us to the other option – the spectre of utility cascades may sometimes cause Bill to lower his effectiveness ratings to a level where he no longer backs initiatives vulnerable to them. But this doesn't vitiate the problem of utility cascades – it capitulates to them. It illustrates the limitations they place upon the deliberations of utilitarian agents. After all, this lowering of efficiency scores is based not on any innate feature of the situation, but on Bill's *own* vulnerability to cascades. One apparently appealing feature of utilitarianism is that it tells us to do the best we can given the limitations of the world as it is – but in this case, one of the limitations of the world is that it contains epistemically rational act-utilitarian agents. If Bill and the other Alt-Fers were not the kind of epistemically rational act-utilitarians they are, things would have gone better.

By the utilitarian's own lights, this is a problem. And it is not anomalous. The preconditions that permit of utility cascades are not rare. First, support for a policy must admit of degree. Second, the effectiveness of a policy must be partly dependent upon its degree of support. Most collective attempts to make the world better – especially political projects – instantiate these features. The two further features which allowed the utility cascade in the climate change case to get *really big* are not rare either. The first was that the efficiency of competing interventions is not independent. This will be the case whenever we must choose between multiple policies addressing the same problem, especially when some are prevention policies, and others mitigations – the more unlikely one policy is to succeed, the more likely the others are to be needed, and hence the higher their expected utility. The second was the presence of multiple informed agents who apportion their support to efficiency. This is also common in political situations. Ironically, it is promoted by the existence of effective altruist organisations.

5. Antecedents

Things would have gone better had Bill either a) not updated his beliefs in the light of new evidence concerning the effectiveness of effactanol and climate change prevention, or b) not changed his degree of support in the light of new beliefs about effectiveness.

Take b) first. It's a familiar claim that sometimes the best outcome in utilitarian terms will be realised if agents do not behave as act-utilitarians. But normally, these arguments assume that people are not perfect utilitarian agents – for example, because their happiness requires such messy things as relationships of partiality or the pursuit of values other than utility-promotion (Parfit 1986: Chapter 1, Railton 1984). Other arguments rely on unusual features of situations, such as the need to thwart a ruthless-yet-rational home invader (Parfit 1986: Chapter 1). No such assumptions are in play in my cases of utility cascades. Indeed, if a utilitarian approach to decision-making were ever apt, one would have thought it would be when donating to charity!

Similar comments apply to the literature on dynamic choice. Familiar dynamic choice problems either involves imperfectly rational agents, that have incommensurable or vague (Tenenbaum and Raffman, 2012) goals, or time-biased (Dougherty 2011) or intransitive preferences (Quinn 1993), or unusual features of the situation, such as the existence of the eccentric billionaire in Kavka's 'toxin puzzle' (Kavka 1983). Utility cascades require no such posits.

With regard to a), the utilitarian argument for epistemic irrationality in the form of 'leaps of faith' goes back as far as Pascal, via William James (1956). But again, these arguments tend to rely either on special facts (such as a punitive God), or on non-rational elements of the scenario (as in James' application of the 'will to believe' argument to cases of romantic love). Again, the argument from utility cascades requires no such elements.

Furthermore, it's often been charged that Pascalian/Jamesian arguments encourage us to do something impossible – to deliberately believe *against* or *beyond* the evidence. This is supposed to presuppose doxastic voluntarism, and, we are told, doxastic voluntarism is false. But ostriching is not believing against or beyond our evidence. It is *avoiding* fresh evidence. Avoiding evidence is not impossible for human beings: much of the world's population appears to manage it regularly.

Finally, the climate change case involved collective action. It's well-known that utilitarian agents can collectively bring about suboptimal results even when acting rationally, in a moral corollary of the prisoner's dilemma (Gibbard 1965). The diagnosis for such collective action problems normally lies in the failure to *co-ordinate*. But Bill's philanthropic friends *do* co-ordinate, in their fashion, through their shared participation in the Alt-F movement. The problem is that there are simply limits to the ways in which act-utilitarians *can* co-ordinate. While they can share information and make plans together, they cannot undertake to perform individually sub-optimal actions. Thus they are prohibited, by their own principles, from falling into a utility cascade once one begins. Indeed, things would have been better had they been *less* co-operative – if Bill had hidden the evidence from the other philanthropists and kept the news of his reduced donations secret.

6. Theoretical Implications

Due to the risk of utility cascades, utilitarianism may frequently call upon utilitarian agents to ostrich, ignoring evidence against the efficacy of interventions they have started to support. What conclusions should we draw from this?

Doody suggests that it is a point against certain non-utilitarian theories that they licence ostrich behaviour. We could say the same of utilitarianism. This would be especially compelling to those who hold that epistemic normativity is somehow prior to moral normativity - that morality *cannot* tell us to shy from evidence.

Another option is for the agent, instead of ostriching, to refuse to change her level of support in the light of changes in the effectiveness of an intervention, instead cleaving to her original decisions even when altering them would be optimific.[4] Rather than being an act-utilitarian, she might become some sort of indirect utilitarian.[5] Thus, the problem of utility cascades may offer another reason to think that it is bad, in utilitarian terms, for agents to adopt an act-utilitarian decision-procedure – that act-utilitarianism is indirectly self-defeating or self-effacing.

The third option is to bite the bullet, and to accept a picture of act-utilitarianism that sometimes licences ostrich behaviour. After all, if what matters most is maximising utility, why shy away from epistemically irrational behaviour when that would, in fact, maximise utility?

7. Practical Implications

Effective Altruism is an increasingly influential social movement inspired by utilitarianism. It is also, as my examples illustrate, doctrinally committed to remaining vulnerable to utility cascades. Why? Because Effective Altruism *combines* the act-utilitarian doctrine that we should apportion our efforts according to efficiency, with a doctrine concerning the *acquisition* of evidence about efficiency: that ‘We should employ the best empirical research methods available in order to determine, as best we can, which efforts promote those values most efficiently’ (Berkey 2018, p147). This is precisely the combination of views which, I have argued, is untenable.

[4] This is to some extent an ethical analogue of McClennen’s (1997) argument for Resoluteness in dynamic choice contexts, although McClennan is only concerned with self-regarding utility considerations.

[5] Perhaps by accepting ‘pattern-based’ reasons for action (Woodard 2019, Chapter 5).

This has important implications for the relationship between utilitarianism and politics. According to the "Institutional Critique", Effective Altruists have wrongly failed to support structural, political or systemic interventions to address the root causes of human misery (e.g. Srinavasan 2015). Effective Altruists have responded that their low valuation of institutional strategies is not mistaken. As Berkey argues, even if structural reforms would have huge effects on utility *were they to work*, they should often be assigned a low efficiency rating on the grounds that *not enough people support them to make them work.*

Berkey argues that an advocate of the Institutional Critique must reject one of two claims. The first is the claim that individuals have very strong reasons to promote increases in welfare by supporting efforts that will do so most effectively; the second is the doctrine quoted above, that agents should proactively seek out the evidence concerning the effectiveness of interventions. He thinks that it is highly implausible – even for a non-utilitarian – to reject either of these claims. He does not even discuss the possibility of rejecting the second. But utility cascades show that a *utilitarian* must reject one of these claims. In particular, since there can be utilitarian reasons to adopt ostrich behaviour, the denial of Berkey's second doctrine hardly seems so implausible.

This connection to the institutional critique should not be surprising, for the Effective Altruist response to structural interventions has something of the form of a utility cascade. Many people who are concerned with world suffering have committed time and effort to securing political reforms. In discouraging people who might otherwise have done so from supporting these initiatives, on the grounds that *not enough* people support them, Effective Altruist organisations *reduce* the effectiveness pursuing of political reform. This in turn reinforces the Effective Altruist case against political engagement.

Probably, the only way to address the root causes of world misery is through structural reforms – the interventions with the highest utility *were they to work* are systemic and political. Whether or not they *do* work is in part dependent on how many people pursue them. But, in a world increasingly influenced by effective altruists, the likelihood of people pursuing these reforms is reduced by arguments that this is an inefficient strategy. Perhaps the world would be better, in utilitarian terms, if Effective Altruists would keep quiet about the difficulty of political reform.

Utilitarianism was born as a radical political movement. Had the early utilitarians sought counsel from today's effective altruists; had they more realistically assessed their prospects of effecting structural changes to their society – of overthrowing centuries of tradition – perhaps they might have directed their attention elsewhere. Instead, they effectively stuck their heads in the sand. They threw themselves into social reform. This motivates a final reflection: perhaps we should give more credit to political reformers who act like ostriches today, and aim to change the world despite the odds.[6]

Bibliography

Berkey, B. 2018. The Institutional Critique of Effective Altruism. *Utilitas* 30 (2):143–171.

Doody, R. Consider the Ostrich: Non-Utilitarians, Ex Ante Interests, and Burying Your Head in the Sand. MS.

Dougherty, T. 2011. On Whether to Prefer Pain to Pass. *Ethics*, 121: 521–37.

Gibbard, Allan. 1965. Rule-Utilitarianism: Merely an Illusory Alternative? *Australasian Journal of Philosophy* 43.

Franzen, J. 2019. What If We Stopped Pretending? *The New Yorker, September 8, 2019,* accessed online at https://www.newyorker.com/culture/cultural-comment/what-if-we-stopped-pretending

Kavka, G. S. 1983. The Toxin Puzzle. *Analysis* 43: 33–6.

McClennen, E. 1997. Pragmatic Rationality and Rules. *Philosophy and Public Affairs* 26(3): 210–58.

Parfit, D. 1986. *Reasons and Persons*. Oxford: Oxford University Press.

Quinn, W. 1993. The Puzzle of the Self-Torturer. In his *Morality and Action*. Cambridge: Cambridge University Press.

Railton, P. 1984. Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs* Vol. 13, No. 2. 134–171.

[6] I would like to thank Simone Gubler, Geoff Sayre-McCord, Jonathan Dancy and two anonymous reviewers for their helpful comments.

James, W. 1956. The Will to Believe. In his *The Will To Believe.* New York: Dover.

Woodard, C. 2019. *Taking Utilitarianism Seriously.* Oxford: Oxford University Press.

Srinivasan, A. 2015. Stop the Robot Apocalypse. *London Review of Books* 37: 3–6 accessed online at http://www.lrb.co.uk/v37/n18/amia-srinivasan/stop-the-robot-apocalypse

Tenenbaum, S. and D. Raffman. 2012. Vague Projects and the Puzzle of the Self-Torturer. *Ethics,* 123: 86–112.