# Two Ways to Satisfy (and No Way to Satisfy Utilitarians)

*Alexandra Zinke, Karlsruhe Institute of Technology, Germany*

## Abstract

Preference utilitarianism holds that an action is morally good iff it maximizes overall preference satisfaction. In principle, there are two ways to satisfy preferences: either you alter the facts such that they fit the subject's preferences, or you change the subject's preferences such that they fit the facts. While standard preference utilitarianism focuses on the first strategy, the present paper will explore the prospects and limits of the second strategy. I will firstly argue that there are cases in which it seems morally right to aim at preference satisfaction by preference change, but secondly acknowledge that an action that induces a global change of preferences doesn't necessarily seem morally right. The real philosophical challenge is to distinguish those cases where altering a subject's preferences is morally right from those where it isn't. The paper ends with a skeptical outlook on the possibility of justifying the distinction on purely preference-utilitarian grounds.

## Introduction

Rationality demands the maximization of one's own welfare. According to utilitarianism, morality demands the maximization of overall welfare. *Preference* utilitarianism subscribes to a desire-fulfillment theory of welfare (also known as *preferentism*): a subject's welfare increases with the fulfillment of her desires / the satisfaction of her preferences. Thus, the core idea of preference utilitarianism can be stated as follows: an action is morally good iff it maximizes overall preference satisfaction. Or, in its prescriptive reformulation: One should choose an action that maximizes overall welfare.[1] Preference utilitarianism will be presupposed throughout this paper.

Let us begin by examining the central notion of preference utilitarianism, *satisfaction (or fulfillment) of preferences*. We will say that a preference is satisfied iff the content of the preference is realized: *S*'s preference that *p* is satisfied iff *p*. The subject need not know about the satisfaction of the preference or experience any feelings of fulfillment. The notion of satisfaction is of course not restricted to the satisfaction of preferences but also applies

[1] Here and in what follows, I use "should", "right", etc. in their moral, not their prudential reading.

to all other pro-attitudes, e.g., to wants, desires, wishes, etc.[2] A pro-attitude is satisfied iff it its matched by the world. Preference utilitarianism thus says that actions should establish a match between the content of a pro-attitude and the world. How can they do so? *Prima facie*, there are two ways to establish this fit: one could change the world such that it fits the pro-attitudes, or one could change the attitudes such that they fit the world.[3]

Usually, preference utilitarianism is read in the manner of world-to-preference direction of fit: we should change the (objective, not preference-related) facts such that the world fits the *actual* preferences of the subjects. If Ann prefers the apple to the banana, we should offer her the apple rather than make her prefer the banana. And, to take a somewhat more serious example, if Ben is starving to death, we should give him food rather than make him want to die. Let me call this reading of the initial utilitarian thesis *world-directed utilitarianism*.[4] More precisely, according to *radical world-directed utilitarianism*, an action is morally good iff it maximizes the overall satisfaction of the *given*, i.e., *actual*, preferences. We call this theory *radical* world-directed utilitarianism as it exclusively values preference satisfaction by changes of the objective facts.

It is important to stress, however, that preference utilitarianism as initially stated is neutral with respect to the two strategies of preference satisfaction: nothing but the overall amount of preference satisfaction counts. Utilitarianism itself is silent about *how* the fulfillment of preferences should be achieved. As John Rawls says, if preference satisfaction is all that matters, then we must be “ready to consider any new convictions and aims, and even to abandon attachments and loyalties, when doing this promises a life of greater overall satisfaction” (Rawls 1982, 181). If all that matters is the amount of preference satisfaction, we could make Ann prefer the banana she already has, instead of supplying her with the factually desired apple. And, instead of giving him food, we can at least try selling to Ben the relief found in finally experiencing the eternal tranquility that only death can yield. Let *radical preference-directed utilitarianism* be the thesis that an action is morally good iff it

[2] Talk of satisfied preferences seems a bit awkward, as preferring appears to be a three-place relation between a subject and two objects: *S* prefers *a* to *b*. However, in this paper I will be a bit sloppy and sometimes use “prefer” as a binary relation (“*S* prefers that *p*”) and sometimes as a three-place, comparative relation (“*S* prefers *a* to *b*”). Furthermore, I will use “*S* desires/wishes/wants that *p*” interchangeably with “*S* prefers that *p*”. Nothing of significance will hinge on this.

[3] Of course there is also a third way: one could combine the two strategies and change both. However, I will here concentrate on the two more conservative strategies of manipulating only one side.

[4] As mental attitudes in general, and preferences in particular, are also ‚parts of the world’, this label is not quite accurate. It is intended to stress the contrast between changing the preferences (i.e. mental entities) themselves and changing the facts at which the preferences are directed (which are often, though not necessarily, non-mental facts).

maximizes overall satisfaction of preferences by changing the preferences such that they fit the actual facts. Again, the theory is *radical* as it exclusively values the generation of preference satisfaction by change of preferences, not by change of objective facts.

Radical preference-directed utilitarianism seems to be a nonstarter. If there is a way to satisfy a given preference (without violating any other preferences), then that action seems morally good – at least from the assumed utilitarian perspective. I will not attempt to defend radical preference-directed utilitarianism. But we can think of a weaker form of preference-directed utilitarianism, *liberal preference utilitarianism*, which allows for both ways of maximizing preference satisfaction: it says that an action is morally good iff it maximizes overall preference satisfaction – independently of whether this is attained by changing facts or preferences. Liberal preference utilitarianism will be the view defended here.

The first section will argue by way of example that radical world-directed utilitarianism is wrong: there are at least some cases in which it seems morally good to change an agent's preferences to ones that are satisfied by the world as it is. The second section addresses some potential problems for preference satisfaction by preference change. It defends liberal preference utilitarianism, but also argues that the theory must be supplemented by a principle that distinguishes cases in which preference satisfaction by preference change is a legitimate option from those where it isn't. The paper ends with the skeptical worry that a distinction of these cases cannot be motivated by purely preference-utilitarian means. Preference utilitarianism thus provides at best an incomplete theory of morally good actions.

# I A Case for Preference-Directed Utilitarianism

I will present two types of cases in which it seems intuitively morally good to establish preference satisfaction by preference change. A note of clarification: There is a huge debate about whether the satisfaction of all preferences or only of the intrinsic ones counts, and about whether actual or ideal preferences are the target. I bracket this discussion as I think that my cases apply also to versions of preference utilitarianism that concentrate on intrinsic and ideal preferences: we should sometimes even change ideal intrinsic preferences.[5]

[5] We lack a precise account of ideal preferences, but the following characterization by Arneson might be helpful: "My ideally considered preferences are those I would have if I were to engage in thoroughgoing deliberation about my preferences with full pertinent information, in a calm mood, while thinking clearly and making no reasoning errors." (Arneson 1989, 83)

## I.a Unrealistic Preferences

Ann, your beloved teenage daughter, deeply desires to become the next big pop star. Unfortunately her voice is terrible rather than terrific. Whenever you listen to her, you become more convinced that her dream will forever remain unfulfilled. What should you do? It is practically impossible for you to change the worldly facts such that your daughter's preferences have a chance of becoming satisfied. No singing lessons will help. The only possible way to make her have fulfilled preferences is by changing them. If you are still aiming at maximizing preference satisfaction, you should try to alter her preferences.[6] You could show her different aims in life, foster her interest in painting or sports so that she will forget about the pop star business, or maybe you should introduce her to punk music.

If it is practically impossible to satisfy a given preference, we should try to reach preference satisfaction by changing preferences. Unrealistic preferences provide the first sample case in which is seems right to change a subject's preferences.[7]

## I.b Conflicting Preferences

Ann has grown up and is now planning her honeymoon. Her true love Ben wants to go to the sea, while she prefers the mountains. Money is sparse, so they cannot do both; love is intense, so they definitely want to go together. They consult you about what to do. What should you do?

Given the circumstances, it is metaphysically impossible to satisfy both Ann's and Ben's preferences. If you are striving for a maximization of preference satisfaction, you should try to change Ann's or Ben's preferences (or both). This will probably be no easy task, but it seems to be the way to go. Only once Ann's and Ben's preferences are in harmony will it be possible to satisfy those of both of them. Conflicting preferences, i.e., preferences that cannot be satisfied simultaneously, provide my second sample case in which it seems morally right to change a subject's preferences.[8]

[6] For reasons of simplicity, we here ignore the preferences of all other moral subjects, e.g. your possible preferences about Ann's preferences.

[7] What should Ann herself do in the above situation? If we follow the above line of reasoning, she should adapt her preferences. See also Bruckner 2009 for a defense of this intuition with respect to a similar case.

[8] Typical cases involving "ill preferences" or "perverse desires" can also be described as cases of conflicting preferences: if Cen desires to torture the cat, Cen's and the cat's preferences are in conflict.

I think that the two presented cases support the view that it is sometimes morally right – at least from a utilitarian perspective – to strive for preference satisfaction by preference change. If that is correct, radical world-directed preference utilitarianism is wrong. As always, however, moral intuitions might diverge. Some readers might have different views on some or all of these cases. Let me observer however, that a defense of radical world-directed utilitarianism requires some justification for the primacy of actual or given preferences over not-yet-actual ones. The core principle that preference satisfaction is of (moral) value has an immediate intuitive appeal that the more sophisticated principle, which exclusively focuses on, and holds fixed, given preferences, lacks. From a purely preference-utilitarian perspective, what should be wrong with adapting preferences to the world – at least sometimes?

In the next section, I will discuss two possible objections to preference satisfaction by preference change. I will reject the first objection, but acknowledge that the second objection points to the limits of preference satisfaction by preference change. We end up with a modest form of liberal preference utilitarianism.

## II Objections to Preference-Directed Utilitarianism

The first objection to preference-directed utilitarianism employs the notion of higher-order preferences, i.e., preferences about one's own preferences. We can think of preferences as ordered in a (possibly infinite) hierarchy. The preferences of order 1 are directed at the world. Preferences of order 1 are, e.g., the preference for an apple, the preference for becoming the next big pop star, or the preference to spend time in the mountains. But the subject will possibly also have preferences that have preferences of order 1 as their contents. For example, the subject might have the second-order preference that her first-order preferences will soon be satisfied, or the second-order preference that no one changes her first-order preference to go to the mountains. Then again, there can be preferences of order 2, etc., *ad infinitum*. In general, preferences of order $n$ + 1 will concern preferences of at most order $n$. (Real agents will often not explicitly entertain many higher-order preferences. However, first, we can also allow for implicit preferences; second, we here consider somewhat idealized agents; and third, and most importantly, we aim at making the *prima facie* objection to preference-directed utilitarianism as strong as possible.)

*Objection (higher-order preferences)*: It is plausible to assume that at least some agents have higher-order preferences that (at least some of) their lower-order preferences are not to be interfered with. Ann wants to become the next big pop star and wants nobody to change that preference of hers. And I have a very strong preference that no neuroscientist changes my preference not to commit suicide today. Thus we can usually not improve overall preference satisfaction by changing preferences of a lower order as this will violate strong higher-order preferences.

*Reply*: This objection to preference-directed utilitarianism applies only to an impoverished version of the theory. Of course, if an agent has relevant higher-order preferences, e.g., the second-order preference *B* not to change the first-order preference *A*, then one should not interfere with *A* in isolation but change *B* first. We must always begin by changing the relevant preferences of the highest order. Thus, before changing Ann's first-order preference to become a pop star, we must change her second-order preference that no one interfere with her first-order pop star preference. (If the hierarchy of preferences is infinite and there is no highest preference to begin with we should change all relevant preferences simultaneously.)[9,10]

Preference satisfaction by preference change, understood correctly, does not violate any higher-order preferences: they are not violated, because they are changed first. However, let me now develop another, more fundamental, objection to preference-directed utilitarianism. It shows that in many cases, realizing preference satisfaction by preference change seems intuitively morally wrong (or at least not morally right).

There is a trivial two-step way to maximize preference satisfaction by preference change: we first delete all unsatisfied preferences – this eliminates any mismatch between preferences and facts – and then generate maximally strong preferences such that they are satisfied by the world as it is, thereby maximizing the overall amount of preference satisfaction. Thus we make the agents maximally desire whatever is actually the case – and only this. If the number of fish in the Amazonas is even, then we should make Ann strongly desire this; if the last dinosaur died on a Tuesday, we should make Ann have a strong preference for

[9] Of course, from a practical perspective this is quite demanding. However, overdemandingness objections seem irrelevant as long as we are discussing only evaluative, not prescriptive, moral principles.

[10] Let me point out a remaining worry: I have assumed that our preferences are ordered in a hierarchy. This excludes the possibility of self-referential preferences like the preference that this very preference not be interfered with, or the very general preference that there be no interference with any preference – including this one. Such self-referential preferences cannot be located at any level in the hierarchy. If the conception of self-referential preferences makes sense and an agent has the preference that there be no interference with this very preference, one cannot change it without violating it. Nevertheless, the negative impact of violating this one preference might be countervailed by the satisfaction of all other preferences, so that even self-referential preferences do not necessarily block maximizing the total amount of preference satisfaction by preference change.

this fact, and so on. I suspect that this seems counterintuitive to many proponents of preference utilitarianism. Let me trigger intuitions a bit with the help of a variant of a well-known thought experiment.

*Objection (the preference adjustment machine)*: Let there be a machine, call it the *preference adjustment machine*, that changes all preferences of an agent such that they fit the facts: it deletes all unsatisfied preferences and creates all preferences that fit the facts. Once an agent is plugged into the machine, she prefers maximally whatever is the case. If everybody is plugged to the machine, the machine creates a brave new world with beings who all maximally desire the same: the world as it is. The result is a maximum total amount of preference satisfaction.[11] Thus, according to liberal preference-directed utilitarianism, we should all plug or be plugged into the machine. Even more: you are morally obliged to plug to the machine, not only yourself, but anybody. Again, this consequence might seem counterintuitive to many. At least, it often does so to me.[12, 13] (If this consequence doesn't seem devastating to you, that's fine! You seem to be a proponent of liberal preference-directed utilitarianism, and the rest of the paper will be of no interest to you.)

*Reply*: I do not want to reject this objection to liberal preference utilitarianism, but rather wish to make precise what exactly it shows. It does not show that we *never* value preference satisfaction by preference change, but it suggests that we *sometimes*, or typically, tend to value preference satisfaction by a change of worldly facts higher than by preference change. If this is correct, then liberal preference-utilitarianism must be supplemented by a principle that distinguishes between cases in which we can maximize preference satisfaction by preference change from cases in which we shouldn't do so. We need a choice principle, or *Preference Principle*, telling us in which cases preference satisfaction by a change of worldly facts is to be preferred over preference satisfaction by preference change. Without such a Preference Principle, liberal preference utilitarianism provides only an incomplete theory of morally good actions.

[11] Of course the total amount of preference satisfaction grows further if the machine additionally creates new bearers of preferences (i.e., new subjects), but let us here concentrate on maximizing the amount of preference satisfaction for already existing beings.

[12] Note that this intuition doesn't fade even given a more restrictive notion of preference change that only allows for changing the content of already existing preferences and does not allow the creation of new preferences.

[13] For a similar, though less radical though experiment, see Parfit 1984, 496: "I am about to make your life go better. I shall inject you with an addictive drug. From now on, you will wake each morning with an extremely strong desire to have another injection of this drug. […] This is no cause for concern, since I shall give you ample supplies of this drug. Every morning, you will be able at once to fulfil this desire."

We cannot here discuss different proposals for a Preference Principle. Yet let me exemplarily introduce one, if only for the purpose of illustration. The following Preference Principle suggests itself:

*Preference Principle*: Satisfy the actual (intrinsic and ideal) preferences first. If impossible (e.g., because of highly unrealistic or contradicting preferences), change the preferences.

The suggested Preference Principle gives priority to the satisfaction of given preferences, but suggests changing the preferences if this is the only plausible possibility leading to preference satisfaction. It thereby captures both the intuition that there are cases in which we can, or should, attain preference satisfaction by preference change, and the intuition that we need not plug ourselves or others into the preference adjustment machine. Thus there seems to be an easy way to complete liberal preference utilitarianism with a suitable Preference Principle.

However, let me stress that the above Preference Principle (and, I fear, most variants of it)[14] does not seem to allow for a justification within preference utilitarianism. From the perspective of preference utilitarianism, there is no reason why we should opt for the satisfaction of actual preferences first. Preference utilitarianism, as stated above, exclusively aims at maximizing welfare, where maximizing welfare is understood as maximizing preference satisfaction. The source of a rationale for the Preference Principle, however, seems to rely on considerations surrounding the “autonomy of the subject” the “subject’s identity”, “oneself being the author of one’s preferences”, or something along these lines. In a purely preference-utilitarian worldview, there is no place – or at least no natural place – for valuing autonomy or the like; the only moral good is proclaimed to be a maximization of satisfied preferences.

# III Conclusion

There are two ways to obtain a satisfied preference: by changing the world such that it fits the actual preferences, or by changing the preferences such that they fit the worldly facts. The common reading of preference utilitarianism focuses on changing the world and leaving the preferences intact. I have defended liberal preference utilitarianism, which allows for

[14] For instance, one could suggest a principle employing the distinction between deliberate and unconscious preference adaption (see, e.g., Elster 1983 and Bovens 1992), or propose a principle referring to Bruckner’s notion of “reflectively endorsed” preference change (Bruckner 2009).

both ways of preference satisfaction. More precisely, I have suggested that there are cases in which preference satisfaction by preference change seems morally right, but also stressed that a universal adjustment of preferences to reality doesn't seem morally right. If you share these intuitions but still want to stick to preference utilitarianism, you must supplement your theory with a Preference Principle that distinguishes these situations. I doubt that such a principle can be justified within a purely preference-utilitarian framework.

# Acknowledgement

I am grateful to Christopher v. Bülow, Wolfgang Freitag, Peter Königs and Christian Seidel for very helpful comments on an earlier version of this paper.

# References

[1] Arneson, Richard J. 1989. "Equality and Equal Opportunity for Welfare." *Philosophical Studies* 56: 77-93.

[2] Bovens, Luc. 1992. "Sour Grapes and Character Planning." *The Journal of Philosophy* 89: 57–78.

[3] Bruckner, Donald W. 2009. "In Defense of Adaptive Preferences." *Philosophical Studies* 142: 307-24.

[4] Elster, Jon. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. New York: Cambridge University Press.

[5] Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

[6] Rawls, John. 1982. "Social Unity and Primary Goods." In *Utilitarianism and Beyond*, edited by Amartya Sen and Bernard Williams, 159-85. Cambridge: Cambridge University Press.