Abstract

Parfit (1997) suggests prioritarianism as a moral theory that incorporates considerations of justice into a broadly utilitarian framework. According to prioritarianism what matters is that we maximize the total sum of (expected) morally weighted utility, where the moral value of utility is *marginally diminishing*. This means that it's more valuable to confer utility to someone who is worse off than to someone better off. Otsuka & Voorhoeve (2009) present a class of examples where the moral verdicts of prioritarianism seem to go awry. They defend their result claiming that adequacy in said examples can only be achieved by prioritarianists if they deny the *moral significance of the separateness of persons*. After briefly introducing the prioritarian framework, Otsukas & Voorhoeves counterexample and argument will be restated. Their objections will be found to be unproblematic for prioritarianists who can simply insist on their moral intuitions concerning the supposed counterexample and the supporting argument will be claimed to be question begging.

Table of contents

1. Prioritarianism

Standard utilitarianism is not sensitive to inequality as a factor that influences the moral status of states of affairs. For this reason Parfit (1997) suggests prioritarianism as a position that keeps the general idea and spirit of utilitarianism but rectifies said deficit. Consider the following two states:

$\mathcal{A}$	200, 100
$\mathcal{B}$	145, 145

In $\mathcal{A}$ every person of one group has 200 units of utility and every person of another *same sized group* has 100 utility. In $\mathcal{B}$ everybody has 145 units of utility. Importantly, the total sum of utilities is greater in $\mathcal{A}$ than in $\mathcal{B}$ but equality in the distribution of utility is greater in $\mathcal{B}$ than in $\mathcal{A}$. Both these features seem relevant when we want to decide which of the two states is to be morally preferred. According to utilitarianism though, we can neglect equality and only consider total (expected) utility to determine the moral status of both options. Thus $\mathcal{A}$ is better than $\mathcal{B}$, because it maximizes the total sum of (expected) utilities; $\mathcal{A}$ sums up to 300 and $\mathcal{B}$ only to 290 and on (expected) average the people in $\mathcal{A}$ have 150 and the people in $\mathcal{B}$ only 145.

Prioritarianism eschews this disregard for egalitarian intuitions. To incorporate equality into moral evaluations Parfit postulates that utility has *diminishing marginal moral value* (cf. Parfit, 1997, p. 213). This means that each additional increment in utility to one person, while not being worthless, still becomes less and less morally valuable. It follows that a benefit in utility is worth more the worse off the beneficiary is. This constitutes a mechanism that prevents ever larger discrepancies in utility from being evaluated as good or better than alternative smaller discrepancies.

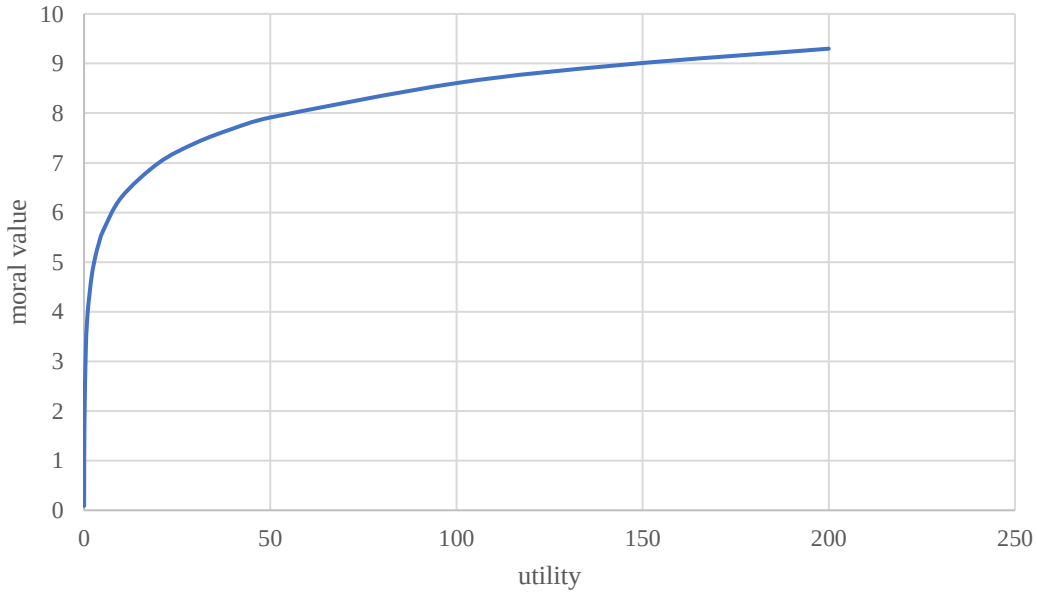


Figure 1; moral value as a function of utility – in this case, for illustration, $ln(x) + 4$.

(Pure) prioritarianism advises then to maximize the total sum of expected morally weighted utility. In the above case of $\mathcal{A}$ and $\mathcal{B}$ we assumed same sized groups in both outcomes, thus we can simply maximize the sum of morally weighted utility. Using the depicted function in *Figure 1* to assign moral value to utility the moral value of $\mathcal{A}$ sums to $\approx 17{,}91$ and the moral value of $\mathcal{B}$ to $\approx 17{,}96$.[1] Narrowly though it may be, $\mathcal{B}$ trumps $\mathcal{A}$ and is the preferred option of prioritarianism. And it needs to be emphasized again: this is not so because prioritarianism thinks of equality as intrinsically valuable but because it uses moral priority of lower utility as a proxy for equality.

One feature of prioritarianism that Parfit insists on (1997, p. 214), is that moral weight/priority is not determined relatively but absolutely. This means that if we have a state $\mathcal{C}$:

$\mathcal{C}$	300, 200

where the discrepancy between the two groups of utility is the same as in $\mathcal{A}$ (200, 100) but the overall utilities involved are higher, the moral weight given to the worse off groups (200 and 100 respectively) is not equal. Although they are equidistant from their counterpart group and the inequality is, in a sense, the same, they are at different absolute utilities, thus the worse off group in $\mathcal{A}$ (100) is given more absolute moral weight than the worse off group in $\mathcal{C}$ (200).[2] Similarly if $\mathcal{A}$ wasn't a state of 200, 100 but of 120, 100 the group of 100 utility would not be given less weight just because there is less inequality, they would still be given the same absolute moral value.

It is primarily this last feature of prioritarianism that will be the object of the following discussion. Prioritarianists do not consider justice to be a comparative matter. It does not matter that some are relatively worse off than others. What matters is only that people are treated as they deserve; that some are absolutely worse off. Contrary to that typical theories of justice, like egalitarianism, take justice to be comparative in nature. If nobody is worse or better off than anybody else than there is no injustice. Otsuka & Voorhoeve challenge the prioritarian account of justice and defend an egalitarian position. We will turn to their critique presently.

2. Otsuka-Voorhoeve counterexamples

With the stage set we can turn to a critique of prioritarianism. Otsuka & Voorhoeve (2009) first present a systematic class of putative counterexamples where prioritarianism seems to not deliver the intuitively – and perhaps empirically – correct moral verdict and second an argument to the conclusion that this

[1] Nothing hinges on the fact that the function for moral value here is $\ln(x) + 4$. It is just being used for illustration. With that in mind: $\ln(100) + 4 \approx 8{,}61$, $\ln(145) + 4 \approx 8{,}98$ and $\ln(200) + 4 \approx 9{,}3$. In general a function for this purpose has to be monotonically increasing and concave, because marginally more utility will never be morally worthless and will be given more weight further down the utility scale than higher up.

[2] If the *moral value* of utility is given by some function $f(x)$, then the *moral weight* is given by the slope, the first derivative $f'(x)$ of that function.

verdict can only be defended if the *moral significance of the separateness of persons* is denied. Thus two questions arise: 1. Is prioritarianisms moral judgement in these examples actually faulty? and 2. How cogent is Otsuka's & Voorhoeve's defense of their counterexample? To begin with *Otsuka-Voorhoeve counterexamples* will be restated. Subsequently we will tackle question 1. and 2.

The *Otsuka-Voorhoeve counterexamples* are pairs of decision problems where: (i) their (rational) decision-theoretic structure is identical as given in *Figure 2* below, (ii) after prioritarian weighting the action with the smaller difference in its payoffs has the greater expected moral value, (iii) the action with the greater difference in its payoffs has the greater expected utility, (iv) for one decision problem the probabilities of consequences are interpreted as: x% of the time everybody affected is at utility v; for the second decision problem they are interpreted as: every time x% of the affected people are at utility v, (v) in both cases an unaffected third party has to make a morally significant decision for the affected people.

		Possible states of the world	
		1	2
Possible actions	a	w, 0.5	z, 0.5
	b	x, 0.5	y, 0.5

Figure 2

(i) Decision problem of an Otsuka-Voorhoeve example: There are two possible actions and two disjunctive and exhaustive states of the world where every state of the world has the same conditional probability – $p(state\ of\ the\ world|action) = 0.5$. w, x, y *and* z *represent utilities of combinations of actions and states of the world –* $u(action\ \&\ state\ of\ the\ world)$ *– and are ordered thus:* $w > x \geq y > z \ \Rightarrow \ w - z > x - y$.

(ii) After prioritarian weighting $f(\cdot)$ *the expected moral value of* b *is greater than the expected moral value of* a*:* $0.5f(w) + 0.5f(z) < 0.5f(x) + 0.5f(y) \ \Rightarrow \ f(w) + f(z) < f(x) + f(y)$.

(iii) The expected utility of a *is greater than that of* b*:* $U(a) = 0.5w + 0.5z > U(b) = 0.5x + 0.5y \ \Rightarrow \ w + z > x + y$.

Some general clarificatory remarks concerning the *Otsuka-Voorhoeve example* are needed. Deviations from the original presentation will be pointed out in the footnotes.

We assume that before a decision is made, everybody affected is at the same level of utility. Only the levels of utility that result from decisions have an influence on the moral judgements. Turning to conditions (i) - (v): Most important about condition (i) is that both decision problems have the same general structure and in particular that the difference between the payoffs of one action (a) are bigger than the difference between the payoffs of the other (b).[3] From condition (ii) it follows that action b has greater expected moral value than a, since the difference between the payoffs of b is less than the difference between the payoffs of a. According to condition (iii) on the other hand a has greater expected

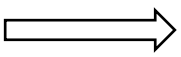
[3] In Otsuka's & Voorhoeve's initial example (2009, pp. 171-173) there still is a difference between the payoffs of what is here called action b. But in a later example (2009, pp. 179-180) they allow for the payoffs to be equal – and thus the difference to be 0. Therefore the condition on the payoffs here is $w > x \geq y > z$.

utility than b, because its payoff difference is greater. Thus, prioritarians would morally favor b over a but rationally they favor a over b.[4]

Condition (iv) is at the heart of the *Otsuka-Voorhoeve example*. It is the only difference between the two decision problems that constitute their counterexample. For one decision the probabilities are interpreted as a fraction of times the decision is or could be made. For our simple case this means: If you choose action a then there is a 50% chance that the world is at state 1 and everybody gets w – the highest possible payoff – and there is a 50% chance that the world is at 2 and everybody gets the lowest possible payoff z. And analogously for action b. For the other decision the probabilities are interpreted not as a fraction of times the decision is made but as a fraction of people affected by the decision. This means: Every time you choose a 50% of the affected population will get w and 50% will get z. And analogously for decision b. Thus, from the first decision always results a population where everybody is at the same level (because beforehand they were at the same level) and from the second decision always results a population where half the population is at one level of utility and the other half at a possibly different level (see *Figure 3*). Thus, in one population everyone is (comparatively) equally well off and in the other population some are (comparatively) worse off than others. Let's call these first decision problems *whole-population decisions* and *split-population decision* the second ones.[5]

Now turning to condition (v). On first appearance it seems obvious. If we are concerned with problematic decisions for a moral theory then the decisions better be morally significant otherwise prioritarianism is not even applicable. But it is not obvious that every decision problem that fulfills the previous conditions is also morally significant. For example, if $w > x \geq y > z > 0$ then it is possible that nobody is harmed by the decision and especially in the case of *whole-population decisions* it is not clear then if it is a moral decision at all.

		Possible states of the world	
		1	2
Possible actions	a	w, 0.5	z, 0.5
	b	x, 0.5	y, 0.5



w w z z
w w or z z
w w z z

50% *of the time everybody at* υ

[4] Otsuka & Voorhoeve first consider the option that both actions have equal expected utility later (2009, p. 178) they also consider the option that $U(a) > U(b)$ and find that it even strengthens their point.

[5] Otsuka & Voorhoeve call *whole-population decisions* '*single-person case*' and *split-population decisions* '*multi-person case*' because they conceive somewhat differently of their example. The first decision problem is one where the morally motivated deciding persons action affects only one person. And the action of the second decision problem affects two people. But the present reconstruction is equivalent with the presentation of Otsuka & Voorhoeve because in both renditions the consequence of the first action is a population where everybody (perhaps just one person) is equally off and the consequence of the second action is a population where half the population (perhaps just one person) is at one level of utility and the other half (the other person) is at a different level. And this is what essentially matters for the example.

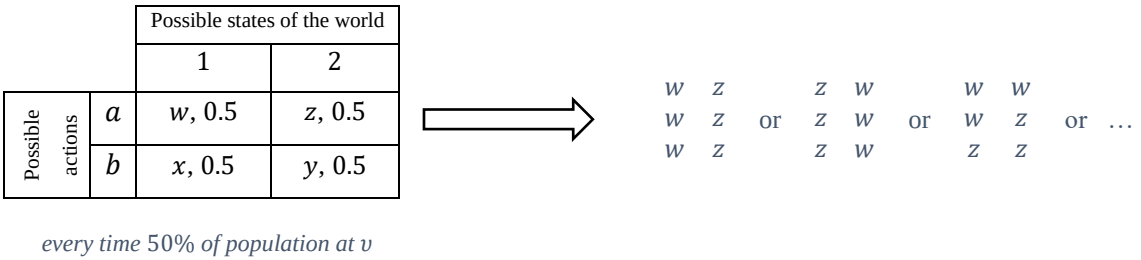


Figure 3

2.1 Intuitions about a shift

Now why do prioritarianism and *Otsuka-Voorhoeve examples* do not get along? The short answer is: Although *whole-population* and *split-population decisions* are structurally identical the difference in the interpretation of probabilities and the resulting difference in the corresponding populations affords a shift in our moral judgement of these cases. Prioritarianism is not able to do justice to this shift because it does not regard justice to be a comparative matter. In *whole-population* cases we deem both option a and b to be morally permissible, Otsuka & Voorhoeve opine. Though a constitutes a gamble since the difference between the payoffs is big, it has the highest expected utility. And if the affected people were to choose for themselves they would choose a. This seems to give us enough moral reason to choose a. On the other hand it does not seem wrong to choose the less riskier option b, which has a higher minimum payoff than a. So choosing b is at least also permissible. But if we change the focus to *split-population* cases, where some people carry the benefits and other people the risks, choosing a becomes much less permissible. In fact Otsuka & Voorhoeve report data that suggests there is an overwhelming consensus that b is the only right action to decide on (2009, p. 174). This putative shift in moral judgement from *whole-population* to *split-population decisions* cannot be explained by prioritarianists; because according to them moral value is not assigned relatively but absolutely and therefore equally in both cases resulting in b being the only course of action for a moral decision maker.

It is important to note that a possible tension only arises for *pure prioritarianism* but not for *pluralist prioritarianism*. Pluralist prioritarianists acknowledge others factors than the moral value of utility as morally relevant. Depending on what those factors are, they might influence the decision in the *whole-population* case differently than in the *split-population* case and justify a shift in moral judgement. The real target of *Otsuka-Voorhoeve examples* is pure prioritarianism. Because according to pure prioritarianism the only thing that matters is maximizing expected moral value which is invariant between the two decision problems (assuming an absolute measure of moral value).

How then can pure prioritarianists deal with this shift in intuitive moral judgement? If they acknowledge that there is this difference between *whole-population* and *split-population decisions*, they would be forced make changes to their theory. The only lever they have to do so is the assignment of moral value. But postulating a difference in moral value of equal utilities would amount to giving up the prioritarian absolute measure of moral weight and value. Introducing a relative measure of moral value

might not be bad in itself, the problem is that the only plausible relative measures in these decisions are relative on the utility of other people. Because the difference between *split-population* and *whole-population decisions* is that in the latter case but not in the former there are other people in the resulting population who are differently off. Pure prioritarianism cannot cite this relation as a reason for differentially assigned moral value (to absolutely equal utility). It would give up too much off its position and become almost indistinguishable from traditional *egalitarian* positions, by conceding justice to be comparative in nature.

Therefore, the only option is denying the alleged shift of moral judgement. There is no quarrel with the moral assessment of *split-population decisions*. On the contrary, prioritarianists agree with everybody else that *b* is the only right thing to do. But with regards to *whole-population decisions* they diverge and have to stomp their feet and claim: Here too we are correct in our moral judgement. *a* simply is not morally permissible and *b* the only possible course of action. Of course just stomping their feet will not do, they must motivate the conflicting opinion somehow. And it seems some such motivation is available for prioritarianism. Because as we have already noted choosing *a* constitutes a gamble. And while there is nothing wrong about gambling with your own life, gambling with the life and well-being of others is not generally permissible. The affected people may be (rationally) risk-affine but perhaps moral decision makers should strive to be risk averse to avoid causing too much harm. Or shorter: Values of utility lower down the scale should be given moral priority; just as prioritarianists postulate. If this justification has some plausibility to it – as I think it does – we arrive at a stalemate. Prioritarianism simply but not implausibly denies the alleged moral difference between the two decision problems of *Otsuka-Voorhoeve counterexamples*. No problem arises because in both decisions *b* is the only right thing to do.

2.2 Separateness of persons

But Otsuka & Voorhoeve do not admit defeat that easily. They give a supposedly independent reason for why their considered moral judgement is more adequate than the prioritarian, why justice is comparative in nature. The reason being the moral significance of the *separateness of persons*. By this they mean:

> In the [split-population] case, there is no single person for whom the prospect of a greater gain is the desirable flip side of exposure to the risk of a lesser loss and for whom the prospect of such gain might be worth the exposure to such risk. […] It follows that rather than simply deciding whether the potential gain outweighs the potential loss to the same person, you must now decide whether the potential gain to the first person outweighs the potential loss to the second person, who would, if this loss materializes, be worse off than the first person. These differences between the [whole-population] and the [split-population] case imbue the potential loss to a person with

> greater negative moral significance in the [split-population] case.[6] (Otsuka & Voorhoeve, 2009, p. 180)

The simple idea is that interpersonal tradeoffs have greater moral significance than intrapersonal tradeoffs. And since *whole-population decisions* only involve intrapersonal tradeoffs but *split-population decisions* involve interpersonal tradeoffs they should be treated morally differently. Thus, according to Otsuka & Voorhoeve, it's the separateness of persons that justifies the shift in moral judgement involved in their counterexample. To support this claim they modify their counterexample and further justify why exactly the separateness of persons is morally significant.

In order to pump more intuitions and drive home their point, Otsuka & Voorhoeve modify the counterexample, specifically condition (ii). It now reads: After prioritarian weighting the action with the greater (instead of the lesser) difference in its payoffs has the marginally greater expected moral value. Payoff w is now big enough to ever so slightly outweigh its own diminishing marginal utility, making a instead of b the morally dominant option for prioritarians. To illustrate: the case of $\mathcal{A}$ and $\mathcal{B}$ at the very beginning instantiates one half of a *standard Otsuka-Voorhoeve counterexample*. It was interpreted as a *split-population decision*. If we increase the highest payoff of $\mathcal{A}$ from 200 to 220 it instead instantiates one half of the *modified Otsuka-Voorhoeven counterexample* because the expected moral value of $\mathcal{A}'$ now is $\approx 18{,}00$ and the expected moral value of $\mathcal{B}$ still $\approx 17{,}96$ – again assuming for illustration $\ln(x) + 4$ to be the prioritarian valuation function. This means that taking the gamble is now not just the best rational but also the best moral decision for prioritarians, both in the *whole-population* and in the *split-population decision*.

Concerning the modified counterexample Otsuka & Voorhoeve again insist that since action $\mathcal{A}'$ only minimally outweighs $\mathcal{B}$ in the *whole-population decision*, the change to the *split-population decision* (and the introduction of the moral significance of the separateness of persons) is enough to tip the scale back in favor of action $\mathcal{B}$. A shift in moral judgement is justified. But again it seems, as argued above, prioritarians do not have to be moved by this modified example alone. They are at liberty to simply deny

[6] They speak of a "greater gain" and a "lesser loss" because they presuppose payoff values where: 1. The payoffs of b are $x = y = 0$, thus if b is chosen nothing changes. 2. The highest payoff w is positive, representing a gain, and the smallest payoff z is negative, representing a loss. 3. The gain of w is greater than the loss of z ($w > |z|$).

They also speak conditionally of the materialization of the risk and the benefit – "if this loss materializes". But in this case the condition for w and z to obtain is not just the choice of action a. In this presupposed instance of their counterexample Otsuka & Voorhoeve introduce an element of chance in the *split-population decision*. If a is chosen "there is a 50% chance that the following will happen": half the population gets w and the other half gets z (2009, p. 180). But this probabilistic element of the consequences of an action actually taking effect is not structurally paralleled in the complementary *whole-population decision* (2009, p. 179). This precludes easy comparison of both decision. What holds for one need not necessarily hold for the other, making the counterexample lose a lot of its potential force. Thus I can only interpret the introduction of this structural dissimilarity as a mistake and take it to be best to omit it from the presentation.

any shift in moral judgement and insist that the extra potential benefit is enough to make $\mathcal{A}'$ morally acceptable even in *split-population cases*.[7]

Prioritarianists don't have to be moved by the modified counterexample. But is the appeal to the moral significance of the separateness of persons able to make them budge? Or to ask differently: Can prioritarianists simply deny that interpersonal tradeoffs have greater moral significance than intrapersonal tradeoffs? They of course can (and must), as long as no decisive reason is given why they cannot. Otsuka & Voorhoeve try to give such reasons but they all beg the question against prioritarianists.

The question begging is most obvious in the quote above. The last sentence points to previously mentioned differences that "imbue the potential loss to a person with greater negative moral significance in the [split-population] case" (Otsuka & Voorhoeve, 2009, p. 180). These differences are: 1. the very fact that tradeoffs in *split-population decisions* are interpersonal and not intrapersonal and 2. in interpersonal tradeoffs some end up worse than others. But that interpersonal tradeoffs are in and off themselves, apart from anything they involve or entail, more morally significant explicitly contradicts the judgements of prioritarianism. As the examples have illustrated prioritarianists do not judge intrapersonal tradeoffs differently than interpersonal tradeoffs. And without further reasons to judge differently they can maintain this position. A part of the population ending up worse than others due to interpersonal tradeoffs may be such further reason for some but definitely not for prioritarianists. It is the bread and butter of their theory that comparative inequality is not intrinsically bad and does not underwrite our moral judgements about justice. Whether prioritarianism is correct in this regard is the very subject of the debate. It is not enough to tell prioritarianists that they are wrong to convince them they are wrong.

But to the credit of Otsuka & Voorhoeve the above quote is not their most serious or even designated attempt to provide independent reason for the moral significance of the separateness of persons. Some pages later they give voice to the actual reason that underwrites their moral judgement:

> In the [split-population decision], one must justify any claim on resources in light of the comparative strength of the claims of others. Those who are relatively worse off have stronger claims to a given increment of improvement simply by virtue of the fact that it is, other things equal, harder to justify improving the situation of someone who is better off rather than someone who is worse off. (Otsuka & Voorhoeve, 2009, pp. 183-184)

[7] A bit of care is indicated: What Otsuka & Voorhoeve do not try to argue is that prioritarianism is at fault assuming that as the extra benefit to the profiting part of the population becomes ever larger that there *is* a point where choosing this extra-large benefit instead of a more egalitarian distribution becomes the morally dominant choice. This is not what they argue because it is a feature of every theory – not just prioritarianism – that recognizes more utility as morally valuable and has to balance the value of additional utility against the value of equality, which will be every plausible moral theory.

But remove "comparative" from the first sentence and "relatively" from the second and you get the prioritarian justification for their moral judgement. Although for them it is the justification for why *b* is the right thing to do in both *whole-population* and *split-population decisions*. People should get what they deserve and those that are (potentially and absolutely) worse off have stronger claims to marginal benefits. That is why (absolutely) higher utilities weigh less than lower utilities. Otsuka & Voorhoeve introduce a relative measure of the moral weight of utility which of course enables a moral differentiation between *whole-population* and *split-population decisions*. But again, whether justice is comparative in nature, whether absolute or relative well-being is what matters for justice is precisely what's at stake in the discussion between prioritarianism and egalitarianism. Therefore, the question is begged against prioritarianists and despite Otsuka's & Voorhoeve's effort they can remain firm believers in non-comparative justice.

3. Conclusion

Otsuka-Voorhoeve examples are exactly the kind of cases where the considered moral judgements of prioritarianism and egalitarianism collide. The differences in the evaluation of these examples bring out at least one key difference of both theories of justice. Egalitarianism predicts a shift in moral judgement because relative inequality as a moral factor is present in *split-population decisions* but not in *whole-population decisions*. Prioritarianism predicts no shift in moral judgement because the amounts of utility and thus of moral weight and value are invariant between both decision problems. To decide which moral theory is superior based on Otsuka-Voorhoeve examples we need independent reasons why one set of moral judgements is better than the other. These reasons can be empirical – like the data concerning *split-population decisions* that Otsuka & Voorhoeve cite – or they can be theoretical – like the levelling-down objection speaks against the intrinsic value of equality (Parfit, 1997, pp. 210-211). But neither can Otsuka & Voorhoeve show that prioritarianism is empirically inadequate with regards to their counterexample nor do they present a theoretical argument that establishes absurd commitments of prioritarianism. The only reasons they provide assume what would need to be shown independently; they beg the question against prioritarianism. But Otsuka-Voorhoeve examples remain especially clear and tractable (empirical) test cases for the theoretical conflict between prioritarianism and egalitarianism. It would probably proof fruitful to submit them to specific empirical testing.

Bibliography

Otsuka, M. & Voorhoeve, A., 2009. Why It Matters That Some Are Worse Off Than Others: An Argument Against The Priority View. *Philosophy & Public Affairs*, 37(2), pp. 171-199.

Parfit, D., 1997. Equality and Priority. *Ratio*, 10(3), pp. 202-221.

Abstract

Transformative decisions are decisions that involve a predictable change in the desires or beliefs of the deciding person. Under the standard theory of rational decision making – *expected utility theory* – this change is commonly represented as a change of the subjective credence or utility function. Chang (2015) takes Ullmann-Margalit (2006) to argue that *expected utility theory* struggles to explain the rationality of *transformative decisions* because it is unclear which credence or utility function, the original or the changed one, must be appealed to. In this essay I argue that the problem for *expected utility theory* is more general and not specific to *transformative decisions*. Smaller, non-transformative decisions and decisions that don't even involve a change of the utility function produce the same problem. The minimal (sufficient) condition for this problem of choice under change to arise is that the decision involves time-inconsistent preferences of the agent. Thus it is also unimportant what causes these time-inconsistencies. To theorize this problem of *expected utility theory* in its entirety it's important to keep its generality in sight.

Table of contents

1. Transformative decisions: Big and small and time-inconsistent

In the philosophical literature 'transformative decisions' are understood as decision that either *directly* change who you are as a deciding person or at least *involve* some personality change of the deciding person (Ullmann-Margalit, 2006, pp. 158-159) (Pettigrew, 2019, pp. 3-5) (Paul, 2014, p. 10). A favorite example in the literature is having a child or remaining childless, where having a child would predictively change the preferences of whoever is deciding. Other examples include big life decisions like: deciding between pension plans and career options or deciding to emigrate or leave your home town. The change in personality is typically modelled as a change in the desires (or beliefs) of an agent and is translated into the classical theory of rational decision making – *expected utility theory* – as a change in the *preferences* (or *credences*) of that agent. That means *transformative decisions* are taken to fulfill at least two conditions:

(i) They involve a predictable change in the preferences of the deciding agent.

(ii) This change is big enough to be personality altering.

There are decisions that satisfy (i) but not (ii); we might call them 'small transformative decisions'. Those are decisions that involve preference-changes which are not big enough to be personality altering. Think, for example, of the decision to purchase any item. Once endowed with the item you predictively come to value it higher than if you didn't purchase it.[1] Thus, purchasing an item will predictively change your preferences concerning that item. Or a different example: Over the course of the day your hunger and appetite will vary. Thus, if you have to decide after breakfast if and what to eat for dinner you have to account for your (predictively) changing preferences concerning food options.[2] Those *small transformative decisions* are seldomly the focus of philosophical investigation and not always rightly so, but more on that later.

Another interesting class of decisions are obtained if condition (i) is strengthened. We can demand the change in preferences to be *time-inconsistent*.[3] This would mean that the change in preferences is such that the preferences before and after the decision are inconsistent with each other. Consider the common phenomenon that people exhibit *time preferences*. They typically give higher value to rewards closer to the present and discount rewards further into the future. 100€ now are better than 100€ in one week. Thus smaller sooner payoffs (SSP) can be equally valuable as larger later payoffs (LLP). Depending on the discount rate, 100€ now can be as valuable as 105€ in one week. Under some such modes of temporal discounting it can happen that SSP is preferred to LLP when the delay between both payoffs

[1] This *endowment effect* is well documented in behavioral economics (Kujal & Smith, 2008), though it is not uncontested (Klass & Zeiler, 2013).

[2] The examples are taken from (Loewenstein, et al., 2003)

[3] (Frederick, et al., 2002) and (Caillaud & Jullien, 2000) give an overview over decisions with time inconsistent preferences in behavioral economic modelling.

occurs closer to the present and that LLP is preferred to SSP when the delay between both payoffs occurs further into the future. It can happen that the preference order between two options reverses; the agent has *time-inconsistent* preferences.[4] Deciding now at t_0 between A at t_1 or B at t_2, where $A < B$, B is preferred over A. But when t_1 arrives A is now (at t_1) preferred over B at t_2. For example: Having to decide between 100€ in one week or 110€ in two weeks, assume people commonly prefer 110€ in two weeks. But having to decide between 100€ now or 110€ in one week – a symmetrical decision in terms of payoffs and length of delay between SSP and LLP but with the difference that the delay occurs closer to the present even involving the possibility of instant gratification – many people prefer 100€ now.[5] Read, et al. (1999) provide another illustrative example: In their experiment 'lowbrow' movies now were preferred over 'highbrow' movies later when people decided for a movie to watch now – with the assumption that 'highbrow' movies are considered more valuable than 'lowbrow' movies. But when deciding for a movie to watch in the future 'highbrow' ones in the farther future where preferred over 'lowbrow' ones in the nearer future. We will see that the decision theoretic reconstruction of such intertemporal decisions involving time preferences that allow for time inconsistencies is interestingly related to the decision theory surrounding (*small*) *transformative decisions*.

A, if not *the* major worry philosophers have concerning *transformative decisions* is if standard theories of rational decision making are applicable to them. In 2. I will present this worry as expounded by Ullmann-Margalit (2006) and argue in 3. that the basis for it is neither strictly condition (i) nor (ii); it is the *strengthened* version of condition (i) that leads to problems. Thus it's decisions involving time inconsistent preferences which are worrisome. This not only includes *transformative decisions* but also *small transformative decisions* and intertemporal decisions with particular temporal discounting dispositions.

2. The limits of expected utility theory

Edna Ullmann-Margalit (2006) like others is concerned with the rationality of 'big decisions'. What she calls 'big decisions' was later termed 'transformative decisions' and we will stick to that. According to Ullmann-Margalit four features characterize transformative decisions (2006, p. 158). Only one of them is important here:

1. It involves a core-affecting change in the desires or beliefs of the agent.[6]

[4] The typical examples of time-inconsistent temporal discounting are *hyperbolic* and *quasi-hyperbolic discounting*. *Exponential discounting* in contrast is time-consistent.

[5] See Frederick et al. (2002) for a survey of experimental research on time preferences. And the Socio-Economic Panel (SOEP) data of the German Institute for Economic Research (DIW) as reported by (Cobb-Clark, et al., 2019) (Cobb-Clark, et al., 2021) for representative population data on time preference dispositions.

[6] More precisely Ullmann-Margalit is only concerned with transformative decisions where the change of preferences is a consequence of the immediate decision and not a product of external factors or the sum of many decisions over time (2006, p. 159). For the same reason I will include *small transformative decisions* and intertemporal

This subsumes condition (i) and (ii) from above. For discussion purposes it is best to keep them separate and translate them into the language of *expected utility theory*, as is commonly done (Pettigrew, 2019, p. 17) (Paul, 2014, p. 22):

(i) It involves a change in the *utility function* (or *subjective probability function*) of the agent.
(ii) This change is big enough to be recognizable as personality altering.

Ullmann-Margalit maintains that transformative decisions delineate a limit of classical rational choice theory in the same way that the atomic level delineates a limit for Newtonian physics. Understanding rationality as maximizing expected utility fails in these cases. This is what she announces to be her point (2006, p. 157) and this is what e.g. Chang takes to be her point (2015, p. 242). But it is not what she ends up arguing for. What she would need to argue to strengthen that point is that we can make transformative decisions rationally but we cannot do it by maximizing expected subjective utility. This would "falsify" expected utility theory. What she actually argues is that in most cases we cannot decide rationally when faced with transformative decisions, rather we have to "take a leap of faith" and pick instead of choose an option (2006, p. 169). In cases where rational choice is impossible rational choice theory is, of course, not applicable but it also doesn't claim to be applicable. Just as it's no problem for a physical theory to not be applicable where no natural laws hold, it is no problem for expected utility theory to not be applicable when no rational choice is possible. Ullmann-Margalit does think that in some cases of transformative decisions we can decide rationally by dividing them into smaller decisions that we are able to decide rationally (2006, pp. 168-169). But the rationality of these smaller decisions can be captured by expected utility theory, thus cases like these also do not speak against expected utility theory. Therefore, what Ullmann-Margalit's arguments try to mark are not the boundaries of *rational choice theory* but those of *rational choice*.[7] She announces a *modus tollens* but delivers a *modus ponens*.

But let's assume that we can in fact choose rationally in cases of transformative decisions and take Ullmann-Margalit's argument, following Chang, to be a modus tollens against the correctness of expected utility theory. The problem she poses then is this: According to expected utility theory, what an agent needs to do to choose rationally in cases of transformative decisions is to maximize expected utility. This means determining the expected utility of every alternative by assigning a utility to the foreseeable relevant consequences of the alternative – this represents the *utility function* of the agent –, weighing those utilities according to how probable she takes the corresponding consequences to be – this represents the *subjective probability* or *credence function* of the agent – and choosing the alternative with the highest sum of weighted utilities. But in the case of transformative decisions this optimization

decisions with certain discounting dispositions in the discussion I will also include decisions with causes for the preference change other than the decision itself, as long as they satisfy *strengthened (i)*.

[7] She even marks these boundaries with the help of rational choice theory as she uses the framework and concepts of expected utility theory to justify her point.

problem is not well-stated because it's unclear from which perspective we should optimize; the perspective of the current self or the perspective of the future self. The modus operandi of expected utility theory is to maximize the total sum of expected utility as given by *the* utility function of the agent and *the* credence function of the agent in the context of the pertinent decision. But the uniqueness conditions of these definite descriptions are not satisfied in the case of transformative decisions. There are multiple – for simplicities sake and without loss of generality we assume two – relevant utility and credence functions of the agent precisely because the agent is transformed in the process of the decision. Since expected utility theory doesn't tell us which utility function should be used to calculate expected utility it cannot guide our rational choice in cases of transformative decisions. Therefore expected utility theory fails to accommodate the rationality of transformative decisions that we have assumed. Let's call this problem '*the problem of choice under change*' (*CUC*).

3. Conditions for the problem of choice under change

Expected utility theory is affected by the problem of choice under change. Importantly though, *CUC* isn't unique to transformative decisions. The minimal conditions – (i) and (ii) – for transformative decisions are neither sufficient nor necessary for it to arise. Expected utility theory suffers from the same problem when we consider *small transformative decisions* and intertemporal decisions of agents with time preferences that allow for time-inconsistent preferences. It is this last feature, time-inconsistent preferences of the deciding agent, that unifies all three mentioned classes of decisions and which we will therefore suggest as a sufficient condition for *CUC*. Let's examine these claims; while arguing first that (i) and (ii) are neither necessary nor sufficient to produce the problem of choice under change we will also demonstrate how the other two types of decisions evoke *CUC*. We'll end by showing how time-inconsistent preferences feature in all three types of decisions and how they suffice for the problem of choice under change to arise.

To show that (ii) isn't necessary for *CUC* we are searching for a type of decisions that satisfies (i) but not (ii) yet still induces the problem of choice under change. We find such decisions in *small transformative decisions*. They can be reconstructed as involving a change in the utility function of the agent – (i) – without this change being big enough to be personality altering – not (ii). Take the decision of buying a new TV: Presently I am not particularly inclined to buy one. But if I were to buy one, despite not wanting to, I would come to regard this as a good decision and wouldn't regret it (remember the *endowment effect*). My changed valuation of the new TV most likely isn't 'core-affecting' enough to constitute a personality change[8] but nonetheless we can ask: which action maximizes my expected

[8] The decision could be framed more radically and in such a way that it plausibly constitutes a personality change. If I presently have an aversion of unknown origin against television but once bought come to find my TV quite nice, other people might say that my purchase transformed me, and be it only in a small way. But for the sake of the argument we consider the scenario in the text.

utility? And answer: This is indetermined as long as it's left unspecified relative to which preferences, present or future, we are optimizing. For the optimization problem to be well-stated we need to know what expected utility function should be maximized. This is unclear as long as it's not given which utility function the expected utility function is referring back on. My present-self utility function values no TV over a TV while my future-self utility function values buying and having a TV over not having a TV. Thus we have shown that condition (ii) isn't necessary to generate *CUC*. Decisions can be small enough to not involve a change in the personality of a person but still cause the problem of choice under change for expected utility theory.

To show that (i) isn't necessary for *CUC* we are searching for decisions that don't satisfy (i) yet still induce the problem of choice under change. We find such cases in intertemporal decisions of agents with utility functions that allow for time-inconsistent preferences. Situations like these need not be modelled as involving two distinct utility functions but can and are typically modelled by invoking one utility function that discounts future payoffs in a particular way. Think again of the example from above: The decision is between 100€ at t_1 – the smaller sooner payoff (SSP) – and 110€ at t_2 – the larger later payoff (LLP). If 100€ at t_1 or 110€ at t_2 are preferred depends on when the delay (between t_1 and t_2) occurs relative to the temporal location of the agent. If you ask the agent at t_1 then she prefers 100€ now (at t_1). If you ask her at t_0 then she prefers 110€ at t_2. Present (t_0) and future-self (t_1) evaluate alternatives using the same utility function yet they diverge with regards to their preferences because the agent discounts payoffs in time with a discount rate that not only depends on the length of the delay of payoff but also on the time the delay occurs relative to her. Now, what maximizes expected utility for an agent in such a situation? Getting SSP or getting LLP? Relative to the preferences of present-self it's LLP relative to the preferences of future-self its SSP. We end up with the problem of choice under change for expected utility theory but without assuming that two distinct utility functions are involved.[9]

But what is the formal reason for this? In the two previous cases the optimization problem wasn't well-stated because it was left unspecified which expected utility function should be maximized (the one using present-self's utility function or the one using future-self's utility function). In this case there is only one utility function and therefore only one expected utility function to be maximized, so this is not the problem. The problem here lies with the discount function (that is integrated into the utility function). To compute a discount factor it requires two time parameters as input: the point in time which is to be discounted and the point in time at which the agent discounts; the former corresponding to the time of payoff and the latter to the time at which the agent is located. In the above example the times of payoff are fixed at t_1 and t_2 respectively for the two decision alternatives. But it is not fixed at which time the agent discounts, at which time she evaluates the consequences of her alternatives. At t_0? At t_1? Or somewhere else entirely? Expected utility theory doesn't mandate what should be the time of

[9] Brocas et al. (2004, p. 51) mention this problem.

evaluation. For this reason the utility function and consequently the expected utility function don't return a single definite value and the optimization problem allows for no definite solution.

One might object that this is only a technical but no practical problem for expected utility theory. We have to make the decision at which time to evaluate the temporally discounted choice alternatives every time we employ a discount function, no matter if it allows time-inconsistent preferences or not. We just make a pragmatic decision about it and continue unobstructedly applying expected utility theory. This reply and strategy indeed works for time-consistent discounting. It does not matter at which time we discount; as long as the length of the delay between SSP and LLP is constant the discount rate between both is constant. Thus either SSP is more valuable than LLP for every time of evaluation or the converse is true. This it what it *means* that discounting behavior is time-consistent and involves no preference reversal. Therefore under time-inconsistent discounting behavior the discount rate between two times varies depending on when the delay occurs relative to the agent, even if the length of the delay is constant. Under such discounting regimes it very much matters *when* choice options are evaluated because their preference order need not be preserved. This is not just a technical problem for expected utility theory, it is a practical problem that hinders applicability. Thus we have shown that condition (i) isn't necessary for the problem of choice under change to arise. Decisions can be free from a change of the utility function of the deciding agent but still cause *CUC* for expected utility theory.

To complete the argument that conditions (i) and (ii) – and therefore *transformative decisions* – are in no special way related to *CUC* we are going to show that (i) and (ii) are together – and therefore also separately – not sufficient for the problem of choice under change to manifest. A witness of this fact will be a decision that involves a change in personality of the agent due to significantly changed preferences as given by the utility function but the rationality of which can still be captured by expected utility theory. An example of such a decision is one where the utility function changes but in a way that hugely amplifies prior preferences of the agent. In this case the optimal choice relative to the present and the future function will be the same. Say, I have to decide whether to become a member in a political party. Presently I lean towards membership. And if I do join I will become an ardent defender of the values the party represents and the most active member the party has ever seen. Friends and family will think I was transformed by my decision. But *CUC* doesn't ensue since both my present and my future self prefer membership over non-membership. Choosing membership maximizes expected utility for all selves; it is the Pareto optimal decision.

One might say that technically we are still left with an ill-stated optimization problem because we still don't know which utility function we should maximize. But this is no practical problem because it doesn't matter which function we optimize since present and future agent (and their utility functions) agree on what the optimal decision is. Again, the problem of choice under change is best understood as posing a practical and not just a technical problem for expected utility theory, because the technical problem alone doesn't hinder application of the theory. And since the given example doesn't evoke a

practical problem for expected utility theory but satisfies both (i) and (ii) we have shown that conditions (i) and (ii) are not sufficient for *CUC*.

This concludes the negative part of the argument. The positive point is still owed. If not (i) and (ii), what is a sufficient condition for *CUC* to manifest?[10] As already said in the first chapter it is the there specified *strengthened condition (i)*: The decision involves time-inconsistent preferences of the agent. Speaking of "*strengthened" (i)* might suggest that the set of decisions satisfying it is a proper subset of the set of decisions satisfying the normal condition (i). This is not the case. *Strengthened (i)* only in one respect strengthens condition (i) but weakens it in another. It strengthens it with respect to actual changes in utility functions. Not just any change suffices. It must be a change that induces a preference reversal, making the preferences of the agent time-inconsistent.[11] *Strengthened (i)* on the other hand weakens condition (i) in that it doesn't mandate a change in utility functions. The utility function of an agent can be stable and remain unchanged as long as it permits time-inconsistent preferences, e.g. when future rewards are discounted in a certain way. A consequence of *strengthened (i)* being a sufficient condition for the problem of choice under change is that it is irrelevant what the cause of the preference reversal is. Be it the passing of time, some cognitive bias, temporal preferences, the very choice itself, external influences or something else entirely; if it causes time-inconsistent preferences it will produce *CUC*.

Decision under time-inconsistent temporal discounting were introduced and defined as satisfying *strengthened (i)*. For (*small*) *transformative decisions* it is equally easy to see that it's the cases where they satisfy the suggested sufficient condition for *CUC* that they become problematic for expected utility theory. Both types of decisions were understood as involving a change in the utility function of the agent. The utility functions represent the preferences of the agent. If they change in such a way that the preference order of the agent reverses then these preferences are time-inconsistent and what maximizes expected utility for one self doesn't maximize expected utility for the other self. *Strengthened (i)* is fulfilled and we get the problem of choice under change. If on the other hand the utility functions change such that the preference order is preserved over time then the preferences are time-consistent and what maximizes expected utility for one self will also maximize expected utility for the other self. *Strengthened (i)* is not fulfilled and we don't get the problem of choice under change. Thus for (*small*) *transformative decisions* it is *strengthened (i)* as well that determines whether they produce *CUC* or not.

4. Conclusion

We have seen that the problem of choice under change is not exclusive to *transformative decisions*. Decision need to satisfy much more minimal conditions – namely: involving time-inconsistent

[10] I only suggest *strengthened (i)* as a sufficient condition. It also might well be a necessary condition but there is room to doubt that since we haven't looked at decisions involving belief reversals which are at least formally able to generate the same problem as preference reversals.

[11] Ullmann-Margalit also makes this amendment later on in passing (2006, p. 167).

preferences – to conjure this problem. These decisions are united by being intertemporal decision that involve different selves of an agent with different preferences where these preferences are inconsistent with each other which entails that there is no Pareto dominant choice option. The option that maximizes expected utility for one self isn't the option that maximizes expected utility for another self. Consequently expected utility theory struggles to accommodate the rationality of such decisions.

The challenge that must be met to defend expected utility theory against this problem is to determine, in a principled way, a single expected utility of every decision alternative. Possible solutions are: (1) a criterion that justifies always prioritizing the utility function at the time of the decision, (2) a criterion that justifies always prioritizing the utility function when the decision takes effect or (3) a function that aggregates preferences into a single preference ordering (as in voting procedures). Different strategies seem appropriate for different types of decisions. And factors like higher-order preferences (for or against change and stability) and commitment devices (like Odysseus uses against the sirens and Parfit's Russian nobleman uses against ageing (viz. Chang, 2015, p. 264)) further complicate the discussion.

Whatever the appropriate normative solution, if we want to theorize the problem of choice under change for expected utility theory it's worthwhile to take not just *transformative decision* but also *small transformative decisions* and intertemporal decision with time-inconsistent preferences into view. The phenomenon is broader than one might think.

Bibliography

Brocas, I., Carrillo, J. D. & Dewatripont, M., 2004. Commitment Devices under Self-Control Problems: An Overview. In: I. Brocas & J. D. Carrillo, eds. *The Psychology of Economic Decisions: Volume 2*. New York: Oxford University Press, pp. 49-65.

Caillaud, B. & Jullien, B., 2000. Modelling time-inconsistent preferences. *European Economic Review,* 44(4-6), pp. 1116-1124.

Chang, R., 2015. Transformative Choices. *Res Philosophica,* 92(2), pp. 237-282.

Cobb-Clark, D., Dahmann, S. C., Kamhöfer, D. A. & Schildberg-Hörisch, H., 2019. Self-Control: Determinants, Life Outcomes and Intergenerational Implications. No 1047, SOEPpapers on Multidisciplinary Panel Data Research, DIW Berlin, The German Socio-Economic Panel (SOEP).

Cobb-Clark, D., Kong, N. & Schildberg-Hörisch, H., 2021. The Stability of Self-Control in a Population Representative Study. *IZA Discussion Papers,* No 14976, Institute of Labor Economics (IZA).

Frederick, S., Loewenstein, G. & O'Donoghue, T., 2002. Time discounting and time preference: A critical review. *Journal of Economic Literature,* 40(2), pp. 351-401.

Klass, G. & Zeiler, K., 2013. Against Endowment Theory: Experimental Economics and Legal Scholarship. *UCLA Law Review 2,* 61(1).

Kujal, P. & Smith, V. L., 2008. The Endowment Effect. In: C. R. Plott & V. L. Smith, eds. *Handbook of Experimental Economics Results.* Amsterdam: North-Holland, pp. 949-955.

Loewenstein, G., O'Donoghue, T. & Rabin, M., 2003. Projection Bias in Predicting Future Utility. *The Quarterly Journal of Economics,* 118(4), pp. 1209-1248.

Paul, L. A., 2014. *Transformative Experience.* Oxford: Oxford University Press.

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

Read, D., Loewenstein, G. & Kalyanaraman, S., 1999. Mixing Virtue and Vice: Combining the Immediacy Effect and the Diversification Heuristic. *Journal of Behavorial Decision Making,* 12(4), pp. 257-273.

Ullmann-Margalit, E., 2006. Big Decisions: Opting, Converting, Drifting. *Royal Institute of Philosophy Supplement,* 81(58), pp. 157-172.

Zimmerman, S. & Ullman, T., 2020. Models of Transformative Decision-Making. In: E. Lambert & J. Schwenkler, eds. *Becoming Someone New: Essays on Transformative Experience, Choice, and Change.* Oxford: Oxford University Press, pp. 73-99.